Nominal Class Assignment in Swahili

A Computational Account

Giada Palmieri[1,*], Konstantinos Kogkalidis[2,1,*]

[1]University of Bologna
[2]Aalto University

Abstract

We discuss the open question of the relation between semantics and nominal class assignment in Swahili. We approach the problem from a computational perspective, aiming first to quantify the extent of this relation, and then to explicate its nature, taking extra care to suppress morphosyntactic confounds. Our results are the first of their kind, providing a quantitative evaluation of the semantic cohesion of each nominal class, as well as a nuanced taxonomic description of its semantic content.

Keywords

Swahili, nominal classification, lexical semantics, computational semantics, topic modeling, unsupervised learning

1. Introduction

Swahili has a grand total of 18 nominal classes (*i.e.*, 'genders'). There is no consensus on the extent to which the assignment of a noun to a given class is determined by its semantic content. We explore this question from a computational angle. Our experiments suggest semantic cohesion among nominal classes, and provide a summary of the taxonomic concepts associated to each class.

2. Background

2.1. Nominal Classes in Swahili

Like other Bantu languages, Swahili has a rich nominal system, where nouns belong to different classes [1, 2], sometimes also referred to as 'genders' [3]. The nominal class is signalled by an affix on the noun itself, and co-referenced with other elements of the sentence through grammatical agreement [4].

In Swahili, verbs require markers that agree with the nominal class of the subject. An example of subject concord is reported below in (1): the noun *mtoto* 'child' bears the prefix of noun class 1 *m-* on the noun, and agrees with the verb through the subject marker *a-*. The same process can be observed in (2) for the noun *mti* 'tree' (class 3), or in (3) for *kitabu* 'book' (class 7).[1]

(1) M-toto a-me-anguk-a.
[1]-child SM[1]-PRF-fall-FV
'The child has fallen.'

(2) M-ti u-me-anguk-a.
[3]-tree SM[3]-PRF-fall-FV
'The tree has fallen.'

(3) Ki-tabu ki-me-anguk-a.
[7]-book SM[7]-PRF-fall-FV
'The book has fallen.'

Table 1 provides an overview of Swahili nominal classes, with their respective nominal affixes and subject concord markers. The division of the nominal classes is based on reconstructions from Proto-Bantu [5, 6, *inter alia*], and it aims at maintaining a correspondence across Bantu languages. Swahili is considered to have a total of 18 nominal classes, but some are missing in standard Swahili (*e.g.*, classes 12, 13 and 18), while others are not uniquely identified by their nominal affix and/or subject concord markers. Odd numbers are traditionally associated with singular classes, and even numbers with plural classes. The first ten classes are in singular/plural pairing relations (*e.g.*, class 2 is the plural form of class 1), while some singular noun classes may lack a plural form or borrow their plural forms from other classes.

There is a long-standing debate on whether Bantu nominal classification is arbitrary [7], or whether it is based on some underlying semantic principles, with specific meanings associated to specific classes [8, 9]. For Swahili, contemporary studies often adopt a stance that lies between these two extremes: nominal classification seems somewhat predictable based on semantic content, though it may often seem arbitrary [2, 10, 1, 11]. This view is also commonly found in textbooks: semantic cues are provided as an aid for the acquisition of Swahili, but accompanied by the admonition that many nouns do not

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Equal contribution. Authorship order was determined through a first-to-five game of rock paper scissors.

giada.palmieri5@unibo.it (G. Palmieri);
kokos.kogkalidis@aalto.fi (K. Kogkalidis)
https://giadapalmieri.github.io/ (G. Palmieri);
https://konstantinoskokos.github.io/ (K. Kogkalidis)

[1]Abbreviations used in the examples: [n] = nominal class; SM = subject marker; PRF = perfect; FV = final vowel.

Table 1
Swahili nominal classes.

Nominal Class	Noun Affix	Subject Concord
1/2	m-/wa-	a-/wa-
3/4	m-/mi-	u-/i-
5/6	(ji-)/ma-	li-/ya-
7/8	ki-/vi-	ki-/vi-
9/10	∅	i-/zi-
11	u-	u-
14	u-	u-
15	ku-	ku-
16	-ni	pa-
17	-ni	ku-

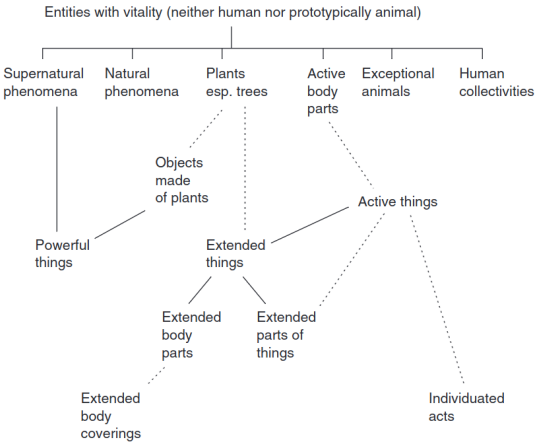
necessarily admit generalizations [12, 13].

Two prominent attempts to examine the semantic categories associated with Swahili nominal classes are provided by Contini-Morava [14] and Moxley [15]. Both studies are cast in a cognitive linguistic framework, and propose networks of meanings and semantic features based on criteria such as resemblance or metaphoric and metonymic extensions. As an example, consider the semantic network for class 3 suggested by Contini-Morava [14] in Figure 1: part of the branching includes the features PLANTS > OBJECTS MADE OF PLANTS > POWERFUL THINGS. Similarly, Moxley [15] suggests a structure of class 3/4 where the notions of 'plants, trees' extends to 'parts of plants' or to objects with 'long, thin, extended shape'. These studies offer valuable insights into the principles underlying nominal classifications, suggesting the potential for more articulate generalizations than are immediately apparent. However, note that they rely on features that were conceived *ad hoc* to account for the categorization of Swahili nouns. Despite this, the nominal classification of several nouns remains unaccounted for [2]. It is unclear whether this is due to features that were overlooked in these studies, or an indication that the classification of some nouns is inherently arbitrary.

2.2. Computational Approaches to Swahili Nominal Classes

Despite the long-standing theoretical debate, computational attempts at semantically characterizing Swahili nominal classes are few and far between. In the context of word sense disambiguation, Ng'ang'a [16] utilizes a collection of manually selected morphosyntactic features in combination with a self-organizing map in order to semantically cluster Swahili nouns. The study finds that including noun prefix features (*i.e.*, nominal class indicators) moderately improves clustering performance, indicating a degree of coherence between semantics and morphology. This improvement is particularly notable for classes 1/2, 7/8, and 11. Olstad [17] trains a naive Bayes classifier over a private, manually annotated dataset that

Figure 1: Contini-Morava's semantic network for class 3.



specifically and explicitly marks the features proposed by Contini-Morava [14]. The approach is framed as an empirical test of Contini-Morava's hypothesis, which the trained model is claimed to experimentally confirm; nonetheless, this assessment is compromised by lukewarm results and a flawed evaluation.[2] More recently, Byamugisha [18] builds a noun class disambiguation system for Runyankore, another Bantu language. The system relies on both a morphological and a semantic component, the latter employing k-NN clustering of word vectors to resolve ambiguities that extend beyond nominal morphology. The work is results-oriented, adopting a task-driven NLP posturing – its only tangible contribution is the system itself.

3. Methodology

Unlike prior works, we are neither interested in preemptively adopting or verifying some existing theory, nor in maximizing discriminative performance metrics in some artificial downstream task. What we *are* interested in is computationally investigating whether semantic content alone is indeed a predictor of nominal class membership. At first glance, word vectors seem to make for a natural starting point. However, language-native word vectors are bound to carry implicit morphological cues, trivializing the mapping to nominal classes (at worst), or obfuscating its semantic aspect (at best). Word vectors (both distributional and predictive) are built on the basis of co-occurrence contexts and/or statistics. The effect of grammatical agreement is that nouns will inadvertently

[2] The key metrics reported are dataset-wide accuracy and per-class area-under-the-curve. Both are over-optimistic: the first tends to favor class-imbalanced datasets, whereas the latter ignores precision and obfuscates the predictive conflict of the competing classifiers.

Figure 2: Example of parsed lexical records.

```
[...
  {"entry": "yahe",
   "definition": "friend, comrade",
   "subject_concord": "a-/wa-"},
  {"entry": "yahe",
   "definition": "commoner",
   "subject_concord": "a-/wa-"},
...]
```

co-occur with verbs that carry subject markers indicative of the noun's class. Case in point, the examples in (1), (2) and (3) contain morphologically distinct entries of the same verbal stem, which disclose the subject's nominal class. The same problem is expounded when using modern segmentation techniques which implicitly account for morphology by incorporating information at the sub-word (*i.e.*, syllable- or character-) level (*cf.* BPE [19], SentencePiece [20], *inter alia*). To bypass the problem, we conduct our analyses on English translations of Swahili nouns. Mediating meaning through a foreign language carries the risk of inducing translation shifts and introducing inaccuracies. That said, we deem it a necessary compromise; the bottleneck completely erases any traces of morphology, which would otherwise confound our results (and their interpretation).

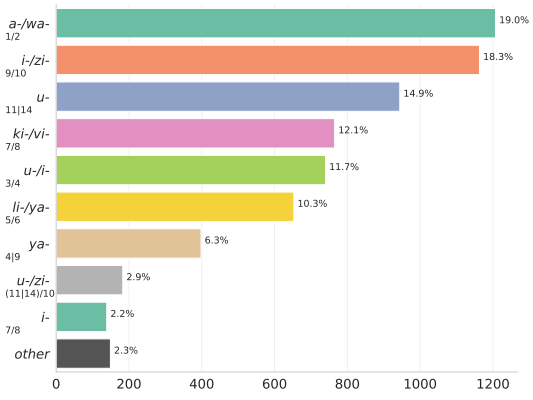
3.1. Data

We first compile a list of nominal lexical entries by consulting the TUKI Swahili-English dictionary. We gather these by scraping the dictionary's online version[3], filtering for pages under the category of Swahili nouns. The scrape yields 5 974 lexical entries. Each lexical entry corresponds to a Swahili nominal homograph. Each homograph is assigned one or more meanings, grouped under one or more subject concord classes. Meanings are provided in English, in the form of (lists of) synonyms, brief descriptions, or mixtures of the two. These are sometimes interlaced with linguistic metadata such as usage examples, apothegms, explanatory comments, *etc.*

The dictionary is consistent in its typographic notation, which allows us to standardize its presentation with a tiny rule-based parser. The parser removes metadata and splits homographs to nominals with unique meanings, gracefully pointing out the occasional inconsistency or error. Guided by the parser, we identify and manually fix common typographic errors. Following our corrections, we are left with a set of 6 341 unique *records*, *i.e.*, triplets of an entry identifier, a meaning and a subject concord class (Figure 2). The distribution of subject concord classes is heavily skewed (Figure 3). We keep records assigned

[3] Available at https://swahili-dictionary.com.

Figure 3: Occurrence counts of subject concord classes.



to one of the 9 most populous classes, which together account for about 98% of the data, and discard the rest. In what follows, we use these subject concord markers as an approximation of the underlying nominal classes.[4] The records we are left with correspond to the nominal classes 1/2, 3/4, 5/6, 7/8, 9/10, 11|14, 4|9 and (11|14)/10; the latter three are necessarily conflated or ambiguous due to their shared morphology.[5]

3.2. Predicting Nominal Classes with a Language Model

Our data allows for a first quantitative inquiry into the semantic uniformity and separation of nominal classes. For our first take, we employ a supervised learning approach. We task a small language model with predicting a record's subject concord class through the phrasal representation of its English definition. The use of a pretrained language model allows the seamless representation of translations that are not strict word-to-word correspondences, promising also the ability to capture subtle semantic distinctions in the process.

We use MiniLMv2 [21], a distilled encoder-only model that has been fine-tuned for sentential similarity using a contrastive learning objective. We apply a 75/25 train/eval split and further fine-tune the model to the task (we follow standard practices, attaching a neural classifier to the model's topmost layer, applied exclusively on the start-of-sequence token). Model selection is based on evaluation loss; we select three models from as many training repetitions over the same split (one model per repetition).

We report means and 95% confidence intervals for the macro- and micro-averaged and per-class F1 scores in

[4] The use of subject concord markers over noun affixes is mandated by the annotation format of the TUKI dictionary.

[5] We use the pipe operator (·|·) to denote disjunction.

Table 2
Macro- and micro-averaged and per-class F1 scores.

M	μ	*a-/wa-*	*i-/zi-*	*u-*	*ki-/vi-*	*u-/i-*	*li-/ya-*	*ya-*	*u-/zi-*	*i-*
34.5±2.6	48.6±1.4	89.4±0.5	35.3±2.9	60.0±3.9	30.2±2.2	42.2±6.2	24.5±4.2	21.4±10.2	5.8±11.4	1.8±5.1

Table 3
Confusion matrix over subject concord predictions.

True \ Predicted	a-/wa-	i-/zi-	u-	ki-/vi-	u-/i-	li-/ya-	ya-	u-/zi-	i-
a-/wa-	299±4	10±0	6±1	9±3	1±1	4±2	0±0	0±0	0±0
i-/zi-	10±2	117±9	43±5	45±13	25±9	26±7	26±7	2±2	1±1
u-	3±0	29±3	153±1	8±2	10±3	9±2	8±4	0±0	0±0
ki-/vi-	13±3	63±8	14±2	57±8	13±3	18±9	5±2	2±2	0±0
u-/i-	1±0	34±2	31±4	22±10	70±13	16±2	8±4	2±2	0±0
li-/ya-	7±0	48±8	13±3	31±10	11±3	34±7	12±5	1±1	0±0
ya-	4±1	39±5	21±3	6±3	4±2	5±4	20±7	0±0	0±0
u-/zi-	1±1	13±2	4±0	9±3	7±2	4±0	2±1	2±2	0±0
i-	3±1	12±2	9±2	4±1	3±2	2±1	2±2	1±0	0±0

Table 2, and per-class predictions in Table 3. Across repetitions, the model is quick to fit the training set, but struggles to generalize, especially on under-represented classes. Despite the fact, performance is significantly better than a probability-weighted random baseline (macro F1 of 14.3).

3.3. Finding the Taxonomies of Nominal Classes with WordNet

Our mixed results paint a nuanced picture. Performance above random affirms that nominal classes are to an extent semantically coherent – even if not *perfectly* so. Performance below perfect, however, offers nothing tangible. The model's shortcomings might be indicative of a semantic dispersion or arbitrariness within nominal classes, but could also be attributed to the model itself, the training process, or the dataset. In either case, we have strong evidence of an (at least partial) overlap between (at least some) semantic and morphological clusters. Other than this confirmation, the supervised approach does not have much else to offer at this stage; over-parameterized black-box models are notoriously hard to extract linguistic insights from. To actually *ascribe* semantic descriptions to nominal classes, we need a better behaved alternative.

For our second take, we employ an unsupervised topic modeling approach. We turn to WordNet [22], a lexical database that maps words to *synsets*: semantically equivalent senses, equipped with periphrastic definitions that are linked together by binary semantic relations.

We begin by matching Swahili records with English WordNet synsets[6]. Matching on a lexical basis is once again impossible; there is no natural correspondence between Swahili nouns and English synsets. As a workaround, we use the same off-the-shelf language model (this time without any additional fine-tuning) to procure semantic representations of Swahili records and English synsets using their respective definitions. We compute a matrix of pairwise scores in the Cartesian product of records and synsets with cosine similarity as our metric. For each Swahili record we then isolate the most similar synsets – no more than 10, and with a similarity score of no less than 0.5. These enact entry points for the Swahili record into the WordNet graph. For each synset, we extract all its *hypernymy paths*: synset sequences that correspond to progressively broader taxonomic generalizations. The *meet* of hypernymy paths originating from multiple synsets associated to a single record correspond to all possible hypernyms of that record. For each record, we weight hypernyms according to their occurrence counts divided by the total number of hypernymy paths in the record; intuitively, hypernyms are assigned a higher weight the more paths pass through them. The process is noisy: error sources include both the matching, and WordNet itself. Nonetheless, we are less interested in the hypernyms of individual records, and more so in their distribution across nominal classes.

On the basis of the above, we have access to the joint probability of nominal classes and hypernyms, $p_{\text{C}\times\text{H}}$, as well as their marginal probabilities, p_{C} and p_{H}. We filter out hypernyms with less than 10 global occurrences, and compute the frequency-weighted[7] pointwise mutual information between classes and hypernyms:

$$\text{wPMI}(c, h) := p_{\text{C}\times\text{H}}(c, h)\ \text{PMI}(c, h) \tag{1}$$

where:

$$\text{PMI}(c, h) := log_2 \left(\frac{p_{\text{C}\times\text{H}}(c, h)}{p_{\text{C}}(c) p_{\text{H}}(h)} \right) \tag{2}$$

Pairs with a positive wPMI score indicate *relevance* (*i.e.*, mutual dependence) between their coordinates – the

[6] A 'native' WordNet would be a better fit for the task, but no mature Swahili version exists as of the time of writing.
[7] The scaling helps alleviate the 'rare event' bias of vanilla PMI.

Table 4
Macro-averaged and per-class weighted relevance between taxonomic descriptors and nominal classes.

a-/wa-	*i-/zi-*	*u-*	*ki-/vi-*	*u-/i-*	*li-/ya-*	*ya-*	*u-/zi-*	*i-*
0.102	0.018	0.040	0.017	0.025	0.016	0.016	0.014	0.009

higher the score, the better a hypernym *describes* a subject concord class. The aggregation of positive scores allows us to quantify and compare the semantic cohesion of subject concord classes given their descriptions – we present these in Table 4. We also present the top 20 extracted descriptors along with their scores in Appendix A. The sum total of positive mutual information between extracted descriptors and subject concord classes under this weighting scheme is approximately 0.26 shannons, suggesting a moderate bidirectional dependency between the two.

4. Analysis

For several classes, our experimental results are congruent with the hypotheses of Contini-Morava [14] and Moxley [15], *inter alia*. Concretely:

- Subject concord class *a-/wa-* is associated with **humans**, **causal agents** and **animacy**; the class is the most semantically coherent and categorically defined; the classifier can accurately predict it, and its taxonomic descriptors are well-pronounced.
- Subject concord class *u-* predominantly refers to **abstract concepts**; the class is the second easiest to predict, and has the most homogeneous description.
- Subject concord class *u-/i-* is mostly associated with **plants**; it is the third easiest class to predict, but predictions are already getting somewhat unreliable.
- Subject concord class *i-/zi-* is semantically **disparate**; its descriptors are heterogeneous and carry relatively low scores. This disparity is consistent with the class' characterization as a 'residual catchall category' [8, 14] where loanwords are often assigned [23]. The only standout descriptor relates the class to **human-made objects**, but the same descriptor dominates also classes *li-/ya-* and *ki-/vi-*.[8] Indeed, the model struggles to tell these three classes apart.

In addition to experimentally affirming existing hypotheses, our approach also yields novel insights and artifacts. With respect to *ya-* and *i-*, the macro-level summary of these two understudied classes reveals an as-of-yet undocumented pattern: both classes lack a singular-plural paradigm, and contain concepts broadly categorized as **abstractions**, albeit of different kinds. This observation may support the correlation between uncountability and abstract meanings noticed in other languages [24, 25]; doing so would however require a thorough examination of these nouns' properties.

From a high-level perspective, we have chosen to isolate the first few highest-ranked semantic components of each class. This ensures backwards compatibility with the literature, but is also a very radical simplification. In reality, our descriptions are fine-grained enough to allow semantically distinguishing between any two classes, even when their primary descriptors overlap. Case in point, *i-/zi-*, *ki-/vi-* and *li-/ya-* have all been reduced to 'human-made objects'; yet the three are actually very different, having only 2 (out of a total of 41) descriptors in common. Moreover, a descriptor is not just a (weighted) concept in isolation, but inherits also the expansive structure of the underlying WordNet it came from. In that sense, our approach does not only describe nominal classes with WordNet synsets, but dually also decorates the WordNet graph with nominal class weights.

5. Conclusions

We explored the relation between semantics and nominal class assignment in Swahili. We approached the question from two complementary computational angles. Verifying first the presence of a relation using supervised learning, we then sought to explicate its nature using unsupervised topic modeling. Starting from a blank slate and without any prior interpretative bias, our methodology rediscovered go-to theories of Swahili nominal classification, while also offering room for further insights and explorations. Our work is among the first to tackle Bantu nominal assignment computationally, and the first to focus exclusively on semantics. Our methodology is typologically unbiased and computationally accessible, allowing for an easy extension to other languages, under the sole requirement of a dictionary. We make our scripts and generated artifacts publicly available at https://github.com/konstantinosKokos/swa-nc.

We leave several directions open to future work. We have experimented with a single dataset, a single model and a single lexical database; varying either of these coordinates and aggregating the results should help debias our findings. We have only looked for semantic generalizations across hyperonymic taxonomies – looking at other kinds of lexical relations might yield different semantic observations. Our chosen metric of relevance is by

[8] Describing *li-/ya* and *ki-/vi-* as human-made objects is in partial alignment with the literature. The two are respectively associated with 'augmentative' and 'dimininutive' meanings [15] and, by extension, with big or small objects [14].

construction limited to first-order pairwise interactions, failing to account for exceptional cases or conditional associations. Finally, we had to resort to computational acrobatics through English in order to access necessary tools and resources. This is yet another reminder of the disparities in the pace of 'progress' of language technology, and a call for the computational inclusion of typologically diverse languages.

6. Acknowledgments

We are grateful to Joost Zwarts and to three anonymous reviewers for their helpful feedback.

References

[1] B. Wald, Swahili and the Bantu languages, in: B. Comrie (Ed.), The major languages of South Asia, the Middle East and Africa, Routledge, London, 2018, pp. 903–924.

[2] F. Katamba, Bantu nominal morphology, in: D. Nurse, G. Philippson (Eds.), The Bantu languages, volume 103, Routledge, London, 2003, p. 120.

[3] P. Spinner, J. A. Thomas, L2 learners' sensitivity to semantic and morphophonological information on Swahili nouns, International Review of Applied Linguistics in Language Teaching 52 (2014) 283–311.

[4] R. M. Dixon, Noun classes, Lingua 21 (1968) 104–125.

[5] A. E. Meeussen, Bantu grammatical reconstructions, Africana linguistica 3 (1967) 79–121.

[6] M. Guthrie, Comparative Bantu, volume 2, Gregg, 1971.

[7] I. Richardson, Linguistic evolution and Bantu noun class system, in: G. Manessy, A. Martinet (Eds.), La Classification Nominale Dans Les Langues Négro-Aaricaines, Centre national de la recherche scientifique, 1967, p. 373–390.

[8] S. Zawawi, Loan words and their effect on the classification of Swahili nominals, Brill Archive, 1979.

[9] J. P. Denny, C. A. Creider, The semantics of noun classes in Proto-Bantu, in: C. G. Craig (Ed.), Noun classes and categorization, John Benjamins Publishing Company, 1986.

[10] M. Krifka, Swahili, in: J. Jacobs, A. von Stechow, W. Sternefeld, T. Vennemann (Eds.), Syntax. An International Handbook of Contemporary Research, De Gruyter, Berlin, 2005, pp. 1397–1418.

[11] L. Marten, Noun Classes and Plurality in Bantu Languages, in: P. C. Hofherr, J. Doetjes (Eds.), The Oxford Handbook of Grammatical Number, Oxford University Press, 2021.

[12] P. M. Wilson, Simplified Swahili, Longman Nairobi; London, 1985.

[13] J. F. Safari, Swahili Made Easy: A Beginner's Complete Course, Mkuki na Nyota; Dar es Salaam, 2012.

[14] E. Contini-Morava, Noun classification in Swahili, Virginia: Publications of the Institute for Advanced Technology in the Humanities, University of Virginia (1994). URL: http://www2.iath.virginia.edu/swahili/swahili.html.

[15] J. L. Moxley, Semantic structure of Swahili noun classes, in: I. Maddieson, T. J. Hinnebusch (Eds.), Language history and linguistic description in Africa, Africa World Press Inc, 1998, pp. 229–238.

[16] W. Ng'ang'a, Word sense disambiguation of Swahili: Extending Swahili language techonology with machine learning, Ph.D. thesis, University of Helsinki, 2005.

[17] J. Olstad, Noun class assignment in Swahili via Bayesan probability, Cambridge Scholars Publishing, 2012, pp. 180–194.

[18] J. Byamugisha, Noun class disambiguation in Runyankore and related languages, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 4350–4359.

[19] P. Gage, A new algorithm for data compression, The C Users Journal 12 (1994) 23–38.

[20] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: E. Blanco, W. Lu (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: https://aclanthology.org/D18-2012. doi:10.18653/v1/D18-2012.

[21] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2140–2151.

[22] G. A. Miller, Wordnet: a lexical database for English, Communications of the ACM 38 (1995) 39–41.

[23] T. C. Schadeberg, Loanwords in Swahili, in: M. Haspelmath, U. Tadmor (Eds.), Loanwords in the world's languages: A comparative handbook, De Gruyter Mouton Berlin, 2009, pp. 76–102.

[24] G. Katz, R. Zamparelli, Quantifying count/mass elasticity, in: Proceedings of the 29th West Coast Conference on Formal Linguistics, 2012.

[25] H. Husić, On abstract nouns and countability, Ph.D. thesis, Ruhr-Universität Bochum, 2020.

A. Appendix

Taxonomic description of nominal classes. Scores are multiplied by $100p_c(c)^{-1}$ to enhance legibility and facilitate direct numerical comparison across classes. Bold face scores indicate higher mutual information. Grayed out descriptors are hyponyms of at least one other descriptor with a higher score.

Subject Concord	Top 20 Descriptors
a-/wa-	person.n.01 (**8.5**), organism.n.01 (**5.8**), living_thing.n.01 (**5.8**), causal_agent.n.01 (**4.1**), physical_entity.n.01 (**3.3**), animal.n.01 (**2.9**), chordate.n.01 (**2.3**), vertebrate.n.01 (**2.3**), whole.n.02 (**2.1**), object.n.01 (**1.6**), bird.n.01 (0.8), aquatic_vertebrate.n.01 (0.7), fish.n.01 (0.7), taxonomic_group.n.01 (0.7), biological_group.n.01 (0.7), adult.n.01 (0.6), bad_person.n.01 (0.6), mammal.n.01 (0.5), unwelcome_person.n.01 (0.5), relative.n.01 (0.5)
i-/zi-	artifact.n.01 (**1.2**), abstraction.n.06 (0.6), instrumentality.n.03 (0.6), matter.n.03 (0.3), device.n.01 (0.3), measure.n.02 (0.3), communication.n.02 (0.3), substance.n.07 (0.2), food.n.01 (0.2), relation.n.01 (0.2), implement.n.01 (0.2), clothing.n.01 (0.2), fundamental_quantity.n.01 (0.1), time_period.n.01 (0.1), color.n.01 (0.1), possession.n.02 (0.1), entity.n.01 (0.1), chromatic_color.n.01 (0.1), substance.n.01 (0.1), visual_property.n.01 (0.1)
u-	abstraction.n.06 (**5.5**), attribute.n.02 (**3.9**), psychological_feature.n.01 (**2.3**), event.n.01 (**1.7**), act.n.02 (**1.5**), state.n.02 (**1.4**), quality.n.01 (**1.4**), entity.n.01 (**1.2**), trait.n.01 (0.7), activity.n.01 (0.7), cognition.n.01 (0.6), property.n.02 (0.6), feeling.n.01 (0.5), condition.n.01 (0.5), group_action.n.01 (0.4), action.n.01 (0.4), change.n.03 (0.3), process.n.02 (0.2), work.n.01 (0.2), immorality.n.01 (0.2)
ki-/vi-	artifact.n.01 (**2.1**), instrumentality.n.03 (**1.1**), object.n.01 (1.0), physical_entity.n.01 (0.9), whole.n.02 (0.7), device.n.01 (0.6), part.n.03 (0.5), thing.n.12 (0.5), body_part.n.01 (0.5), structure.n.01 (0.3), symptom.n.01 (0.2), evidence.n.01 (0.2), container.n.01 (0.2), covering.n.02 (0.2), information.n.02 (0.2), implement.n.01 (0.2), communication.n.02 (0.2), clothing.n.01 (0.2), relation.n.01 (0.2), location.n.01 (0.2)
u-/i-	plant.n.02 (**2.6**), vascular_plant.n.01 (**2.6**), woody_plant.n.01 (**2.0**), tree.n.01 (**1.6**), event.n.01 (0.7), happening.n.01 (0.5), whole.n.02 (0.5), dicot_genus.n.01 (0.5), object.n.01 (0.5), angiospermous_tree.n.01 (0.4), psychological_feature.n.01 (0.4), wood.n.01 (0.4), plant_material.n.01 (0.4), herb.n.01 (0.4), shrub.n.01 (0.3), sound.n.04 (0.3), action.n.01 (0.3), change.n.03 (0.3), material.n.01 (0.3), act.n.02 (0.3)
li-/ya-	artifact.n.01 (**1.5**), object.n.01 (0.9), physical_entity.n.01 (0.8), instrumentality.n.03 (0.7), whole.n.02 (0.5), thing.n.12 (0.5), part.n.03 (0.4), matter.n.03 (0.4), body_part.n.01 (0.4), structure.n.01 (0.4), natural_object.n.01 (0.3), container.n.01 (0.3), edible_fruit.n.01 (0.2), solid.n.01 (0.2), food.n.02 (0.2), plant_organ.n.01 (0.2), plant_part.n.01 (0.2), reproductive_structure.n.01 (0.2), shape.n.02 (0.2), substance.n.01 (0.2)
ya-	abstraction.n.06 (**3.3**), psychological_feature.n.01 (**2.0**), event.n.01 (**1.6**), act.n.02 (**1.3**), entity.n.01 (0.9), attribute.n.02 (0.7), speech_act.n.01 (0.7), matter.n.03 (0.6), state.n.02 (0.6), relation.n.01 (0.5), group_action.n.01 (0.5), communication.n.02 (0.4), cognition.n.01 (0.4), substance.n.01 (0.3), phenomenon.n.01 (0.3), process.n.06 (0.3), natural_phenomenon.n.01 (0.3), activity.n.01 (0.3), feeling.n.01 (0.3), request.n.02 (0.3)
u-/zi-	artifact.n.01 (**3.3**), object.n.01 (**2.9**), physical_entity.n.01 (**2.4**), whole.n.02 (**1.9**), thing.n.12 (**1.4**), part.n.03 (**1.4**), body_part.n.01 (**1.2**), instrumentality.n.03 (**1.2**), implement.n.01 (0.7), palm.n.03 (0.6), part.n.02 (0.6), location.n.01 (0.5), natural_object.n.01 (0.5), device.n.01 (0.5), body_covering.n.01 (0.5), indefinite_quantity.n.01 (0.5), hair.n.01 (0.5), decoration.n.01 (0.4), poem.n.01 (0.4), appendage.n.03 (0.4)
i-	abstraction.n.06 (**3.2**), region.n.03 (**1.0**), location.n.01 (1.0), psychological_feature.n.01 (0.9), matter.n.03 (0.7), cognition.n.01 (0.6), attribute.n.02 (0.6), entity.n.01 (0.6), substance.n.01 (0.6), district.n.01 (0.5), substance.n.07 (0.5), administrative_district.n.01 (0.5), gathering.n.01 (0.5), relation.n.01 (0.5), state.n.02 (0.5), geographical_area.n.01 (0.5), group.n.01 (0.5), condition.n.01 (0.5), process.n.06 (0.5), physical_phenomenon.n.01 (0.5)