MGEN: Millions of Naturally Occurring Generics in Context

Gustavo Cilleruelo Calderón **Emily Allaway** **Barry Haddow** **Alexandra Birch**
School of Informatics, University of Edinburgh
g.cilleruelo-calderon@sms.ed.ac.uk {emily.allaway, bhaddow, a.birch}@ed.ac.uk

Abstract

MGEN is a dataset of over 4 million naturally occurring generic and quantified sentences extracted from diverse textual sources. Sentences in the dataset have long context documents, corresponding to websites and academic papers, and cover 11 different quantifiers. We analyze at scale the features of generic sentences, with interesting insights: generics can be long sentences (averaging over 16 words) and speakers often use them to express generalisations about people.

MGEN is the biggest and most diverse dataset of naturally occurring generic sentences, opening the door to large-scale computational research on genericity. It is publicly available at gustavocilleruelo.com/mgen.

1 Introduction

Generics are sentences that express generalisations without making use of explicit quantifiers. Examples of generics are *ravens are black* or *ticks carry lyme disease*.

Several features of generics make them difficult to account for semantically (Carlson and Pelletier, 1995): they are permissive to exceptions (*ravens are black* is acceptable even if albino ravens exist) and the quantifications they convey have paradoxical dynamics (Leslie, 2008). If we paraphrase the previous generics as explicitly quantified, we would have *most ravens are black* but *few ticks carry lyme disease*: the same linguistic structure conveys generalisations at opposite ends of the quantification spectrum.

In this work, we introduce MGEN, a dataset designed to provide a solid foundation for research on generic sentences in English. MGEN has 4.1 million samples, with over 3 million generics and 1 million explicitly quantified sentences with 11 different quantifiers. All sentences are naturally occurring and include the context document in which they originally appear.

To motivate the design of MGEN, we conduct an extensive review of datasets of generic sentences and argue that existing datasets have many shortfalls: they are either small, rely on synthetic samples or have no context, despite theoretical works showing the importance of context for the semantics of generics (Sterken, 2015; Almotahari, 2023).

In order to mine generic sentences from massive corpora, we introduce a two-step pipeline: a syntactic filter detects bare plurals (this is the most common syntax of the subject for generics, see §2) with the required verb features and then a binary classifier labels them as generic or not. We apply this pipeline to a subset of the ZYDA (Tokpanov et al., 2024) dataset (a language model pre-training corpus) to collect a diverse and accurate (as per human annotators) dataset of generic sentences.

We analyze the corpus-level characteristics of MGEN and find that its generic sentences are longer than those usually considered in the literature, where running examples are much shorter than the average 16.65 words in our dataset. Analysing the word frequencies of our dataset, we find that speakers use generics most often to generalize about *people*.

Our contributions are: (i) MGEN, the largest dataset of naturally occurring generics in context, (ii) a pipeline for the extraction bare plural generics from textual sources, (iii) a review of existing datasets of generics and (iv) a preliminary corpus-level analysis of the characteristics of generic sentences.

2 Background: generics & quantifiers

Generics have kind terms in their subject position (i.e. words or phrases used to categorize or label groups of entities) and their verbs are inflected for third person plural present indicative. They are used either to make claims about those kinds (*dinosaurs are extinct*) or to attribute properties to

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 244-260.
Eugene, Oregon, July 18-20, 2025

Source	Sentence
RefinedWeb	Soybeans contain an inhibitor of trypsin, an enzyme important for digestion, but it can be destroyed by cooking.
SlimPajama	Cucumbers are high in an antioxidant called beta-carotene, which your body turns into vitamin A. May ease muscle cramps.
The Pile	Starving people grab the bread first and run with it.
arXiv	Colexification networks encode affective meaning.
peS2o	Car seats save lives.

Table 1: Examples of generic sentences from the different sources of MGEN. More examples in Appendix F.

individuals in those kinds (*beetles have protective wing covers*).

Following most of the linguistics and philosophy of language literature, we consider only *bare plural* generics (Carlson and Pelletier, 1995; Leslie, 2007a). Bare plurals have noun phrases in plural form without a definite or indefinite article[1]. Throughout the paper, we will use *bare plural sentence* to refer to sentences with the syntax of a bare plural generic (i.e. with the same inflection of the verb), even if those sentences are not generics.

The standard view in linguistics is that generics are quantificational: there is an unpronounced operator GEN that takes a role similar to adverbial quantifiers in the logical form of the sentence (Lewis, 1975; Carlson, 1977b; Carlson and Pelletier, 1995; Cohen, 1999b; Kirkpatrick, 2024).

In contrast, recent influential accounts of generics have been non-quantificational: Leslie (2008) gives generics the privileged role of expressing default or primitive generalisations, Sterken (2015) argues that quantification cannot capture the full context-sensitivity of generics and Nickel (2016) relates generics to a notion of normality grounded in explanatory considerations rather than the prevalence of the property in the kind.

The rich landscape of theories of generics, as well as their far-reaching implications into fundamental aspects of human cognition, has made the study of generic sentences a highly debated topic in recent years (e.g., Cohen, 1999a; Tessler and Goodman, 2016; Stovall, 2019; Nguyen, 2020; Bosse, 2021; Almotahari, 2022; Kirkpatrick, 2023; Neufeld et al., 2025)

In the field of natural language processing, recent works study how language models deal with aspects of genericity such as exceptions, property inheritance (Allaway et al., 2024) and quantification (Ralethe and Buys, 2022; Collacciani et al., 2024). Cilleruelo et al. (2025) uses language models to study the semantics of generic sentences, such as their implicit quantification.

3 Related work: datasets of generics

Several datasets exist that specifically target generics. We compare these datasets across four dimensions (Table 2): total samples, quantified sentences, context and origin (natural or synthetic).

We consider *natural* sentences to be only those that have been extracted from human-written sources and *synthetic* those have been either generated by language models, built with rule-based methods or constructed by researchers or annotators. We also include quantified sentences as a requirement for datasets of generics as these are a key contrast class. Similarly, context plays an important role on the semantics of generics.

GENERICSKB (Bhakthavatsalam et al., 2020) is a dataset that is composed of both naturally occurring generic and quantified sentences in context and synthetic examples derived from knowledge bases.

To source the naturally occurring samples, $3.5M$ candidate sentences are extracted from different corpora (Wikipedia, ARC and Waterloo) through 27 hand-crafted lexico-semantic rules. A subset of those are manually annotated and used to train a BERT-based binary classifier (generic and not generic).

This classifier is used to score the $3.5M$ candidate sentences to curate GENERICSKB-BEST: a collection of the best-scoring naturally occurring sentences ($N = 774,621$) augmented with synthetic generics derived from knowledge bases ($N = 246,247$). Some sentences are quantified with *all*, *most*, *some*, *many*, *every*, *much*, *more*, *often*, *usually*, *always*, *sometimes*, *frequently*.

[1] *Tigers have stripes* is a bare plural generic, which can also be expressed in English with the definite (*the tiger has stripes*) or indefinite (*a tiger has stripes*) articles.

Dataset	Scale	Quantifiers	Context	Sources
MGEN (Ours)	$4.1M$	Yes (11)	Yes	Natural (ZYDA)
GENERICSKB-BEST (Bhakthavatsalam et al., 2020)	$1M$	Yes (13)	Yes	Natural (Waterloo, SimpleWiki, ARC) Synthetic (WordNet, ConceptNet, TupleKB)
CONGEN (Cilleruelo et al., 2025)	2872	Yes (3)	Yes	Natural (DOLMA)
GEN-A-TOMIC (Bhagavatula et al., 2023)	$> 8M$	Yes (3)	No	Synthetic (GPT2-XL with I2D2)
Animal generics (Ralethe and Buys, 2022)	75,002	No	No	Mixed (GENERICSKB)
EXEMPLARS (generics) (Allaway et al., 2024)	$16,655$	No	No	Mixed (GEN-A-TOMIC, Animal generics)
Dataset in (Collacciani et al., 2024)	1837	Yes (5)	No	Synthetic (human annotations)
Norwegian generics (Kurek-Przybilski and Adam, 2022)	170	No	Yes	Natural (encyclopedia entries)

Table 2: Comparison between existing datasets of generic sentences. MGEN is comparable in size with synthetic datasets but is comprised of naturally occurring sentences in context.

Cilleruelo et al. (2025) introduce CONGEN, a collection of 2873 naturally occurring generic and quantified sentences in context. Because the dataset is manually curated, it is small and only contains data for 3 quantifiers (*all*, *most* and *some*).

The biggest dataset of synthetic generics is the GEN-A-TOMIC corpus (Bhagavatula et al., 2023). Sentences in GEN-A-TOMIC are generated by GPT2-XL (Radford et al., 2019) through knowledge distillation with self-imitation algorithm. Although GEN-A-TOMIC has over 8 million utterances, because they are generated with a small language model, these are not in context and the only quantifiers included are *generally*, *typically* and *usually*.

Ralethe and Buys (2022) select generics and quantified sentences from GENERICSKB by filtering for animals, curating a subset of 75,002 generics. This collection of animal generics is combined with examples from GEN-A-TOMIC to create datasets of synthetic generics exemplars (i.e. cases where the generic does and does not hold) (Allaway et al., 2023, 2024), which contain generic sentences, as well as their derived exemplars.

To conduct experiments on language models, Collacciani et al. (2024) collect 1873 sentences from three sources, all crafted either by researchers or annotators (Herbelot and Vecchi, 2016; Urbach and Kutas, 2010; Misra et al., 2023). Sentences in this dataset are extremely short (average length is 3.73 ± 1.03, median is 3) and all are annotated with a quantifier (*all*, *most*, *some*, *few*, *no*).

All datasets considered so far, as well as MGEN, are in English. In Norweigan, Kurek-Przybilski and Adam (2022) manually extract 170 generics in context from encyclopedic texts.

Table 2 compares the reviewed datasets of generic sentences in terms of total samples, inclusion of quantified sentences, context for the utterances and data origin. Our dataset, MGEN, has the scale of GENERICSKB and GEN-A-TOMIC, but without the need of synthetic examples (whether generated or constructed from knowledge bases) and includes context documents for all generic as well as quantified utterances.

4 Methodology

This section details the construction of the MGEN dataset. We first describe the high-level objectives for the creation of the dataset, based on the generics literature and the shortcomings of existing datasets. Then, we detail the extraction of generics and quantified sentences at scale from a large corpus by leveraging syntactic (§4.4) and semantic (§4.5) characteristics of generics.

4.1 Design choices

MGEN is built to include a massive, diverse amount of naturally occurring generic sentences with their respective contexts. In this section we go over the principles that guide the construction of the dataset.

Naturally occurring. We focus on naturally occurring generic sentences, as it would be hard to assess the acceptability of synthetic samples without assuming a theory of generics or conducting extensive human annotation studies, since the semantics of generics are not well understood (§2).

Context. Many works argue that the context radically affects what generic sentences express, for example, in terms of both quantificational strength and flavor (Sterken, 2015; Almotahari, 2023). To mine generic sentences, we choose a corpus structured in documents (more details in §4.2) and keep the full context document of each sample.

Bare plurals. We focus on generics that are bare plurals (§2) and only at the beginning of a sentence. This makes detection at scale more tractable, by, for example, omitting nested generics in *that* clauses (e.g. *she maintains that the belief that technology improves education is widely accepted*).

Quantifiers. Generics and quantified sentences are closely related, as both are used to express generalisations. We collect quantified sentences with the following structures: *quantifier + bare plural sentence*, *bare plural noun phrase + quantifier + verb* or *bare plural noun phrase + verb + quantifier*. We consider the following 11 quantifiers: *all*, *most*, *many*, *some*, *few*, *no*, *often*, *generally*, *typically*, *usually*, *normally*.

4.2 Data sources

Training language models requires large collections of clean textual data, which can also be used for data mining. We use ZYDA (Tokpanov et al., 2024), an open-source dataset built by collecting text from different high-quality sources and performing uniform filtering and deduplication. We run our generic extraction pipeline on the following components of ZYDA (Appendix E; Table E.3): RefinedWeb (Penedo et al., 2023), SlimPajama (Soboleva et al., 2023), the Pile (Gao et al., 2021), peS2o (Soldaini and Lo, 2023) and arXiv (Kenney, 2023).

RefinedWeb, SlimPajama and The Pile primarily consist of data scraped from the web, while the much smaller peS2o and arXiv are composed of academic publications.

4.3 Generic sentence extraction

ZYDA is structured in documents: roughly the text in a website, a scientific article or similar. Each document is first split into sentences (`blingfire`[2]). Then, a lightweight syntactic filtering step selects sentences where either (i) the first word is one of the quantifiers of interest, or (ii) there is a *plural noun* in the first 4 words of the sentence (`flair` (Akbik et al., 2019)).

These candidates are then run through two filtering steps: a syntactic one that ensures these are bare plurals with verbs inflected for third person present indicative and a semantic one, that filters for sentences that express generalizations. This latter step is necessary as the bare plural generic syntactic construction can also have existential readings, where the subject refers to specific instances instead of to a kind in general, e.g. *tigers are in the front lawn* or *blue arrows indicate acceleration* (also see Appendix F; Table F.6).

We detail the construction of each filtering step in §4.4 and §4.5 respectively.

4.4 Syntactic filtering (bare plurals)

The syntactic filtering step in the pipeline receives candidate sentences with plural nouns in the early words and performs a more in-depth dependency analysis to select only bare plural sentences.

The part-of-speech and dependency parsing of the sentence is conducted with the `stanza` python library (Qi et al., 2020). After parsing the sentences, we keep those that meet the following three conditions:

1. The nominal subject is a plural noun or a plural proper noun (`nsubj` or `nsubj:pass` in the case of passives).

2. The root of the nominal subject is a verb or an auxiliary (`VERB` or `AUX`). If there is a copula (`cop`) or a passive (`aux:pass`), take that as the verb.

3. The verb has present tense, indicative mood, plural number and third person.

4.5 Semantic filtering (genericity)

The syntactic filtering step yields bare plural candidate sentences, but these include noisy and non-generic samples. To get high quality generics from these candidates, we apply a further step in which a binary classifier scores whether the bare plurals are generic or not.

This classifier is designed to filter out: (i) sentences that although they may contain a generic it is not at the beginning[3], (ii) sentences that are

[2] https://github.com/microsoft/BlingFire

[3] A common occurrence are titles of paragraphs or sections that get parsed at the beginning of the sentence, for example: *Gaussian Mixture Models Gaussian mixture models are*

ungrammatical or noisy and (*iii*) bare plurals that have existential (non-generic) readings (Table F.6).

We use a ROBERTA model (Liu et al., 2019) as the architecture for the classifer, which we train on a small collection of generics and non-generic bare plurals. The generics are sampled from GENERICSKB-BEST and the non-generics are generated by GPT-4 (OpenAI et al., 2024), by iteratively finding missclassified examples to make the training data more robust. The classifier achieves over 0.97 F-1 score in a test set based on CONGEN and synthetic non-generic bare plurals. More details on classifier training and evaluation are found in Appendix A.

In the case of sentences that start with a quantifier, which are not bare plurals and are outside of the training distribution of the generics classifier, we remove the quantifier word and calculate the score of the resulting bare plural. This ensures that we pick out quantified sentences that are comparable to generics in terms of being generalizations as opposed to existential. We want to keep in the dataset sentences like *all tigers have stripes* but not *all tigers in the cage are male*.

Some quantified sentences begin with a bare plural rather than a quantifier (e.g. *tigers are normally striped*). For these sentences, we check if there is an adverbial quantifier that has as syntactic head the root of the sentence, and label them with the corresponding quantifier (if the quantifier is not in the main clause, the sentence is labeled as generic).

We include sentences that receive a genericity classifier score 0.8 or greater for the MGEN dataset. This value is chosen by manual inspection of the data. The full unfiltered bare plurals data is also made publicly available.

5 MGEN: Statistics & Analysis

In this section we summarize the statistics of the MGEN dataset (§5.1) and present two quality analyses: human annotation to asses the genericity of the collected sentences (§5.2) and a comparison in terms of diversity with existing datasets (§5.3).

5.1 Statistics

We mine generics from a total of $50,534,844$ ZYDA documents (23% of the corpus). After the syntactic filtering of sentences for bare plurals, we end up with $16,771,049$ sentences, of which $\mathbf{4,146,760}$ make up the final MGEN dataset after receiving a score of 0.8 or higher by the generics classifier.

formed by combining multivariate normal .... Note how the title (*Gaussian Mixture Models*) makes it so that the generic is not at the beginning.

	Candidates	**Generalizations**
GEN	$14,303,840$	$3,183,293$
All	$502,629$	$82,752$
Most	$332,698$	$173,021$
Many	$389,606$	$188,419$
Some	$547,308$	$225,171$
Few	$22,164$	$8,085$
No	$47,146$	$4,121$
Generally	$116,901$	$53,015$
Typically	$124,522$	$53,046$
Often	$253,306$	$107,926$
Usually	$138,207$	$59,148$
Normally	$19,969$	$8,763$
TOTAL	$16,771,049$	$4,146,760$

Table 3: Number of generics and quantified sentences after syntactic (candidates) and semantic (generalizations) filtering during the construction of MGEN.

Source composition. The final dataset contains over 3 million sentences from internet crawls (RefinedWeb, The Pile and SlimPajama) and around 1 million sentences from academic sources, peS2o and arXiv (Appendix E; Table E.4). Of the total 4.1 million samples, about 3 million are bare plural generics, while the rest is made up of the 11 quantifiers in different proportions (Table 3).

Context documents. For every sentence in MGEN, we include the document from ZYDA that contains it. These documents correspond to websites or papers and are generally long, averaging over 5000 words. For comparison, the context documents in the samples of GENERICSKB-BEST are much shorter, with an average of 147 words.

Sentence length. We compute the length of sentences in words by splitting sequences by whitespaces. Figure 1 compares sentence length distributions for the naturally occurring examples in GENERICSKB-BEST, the generic (not quantified) sentences in MGEN and the lengths in a sample of 20,000 context documents from MGEN (Figure 1).

Generic sentences in MGEN have an average of 16.65 ± 8.2 words and a median of 15 words: generics are often long sentences. Although generics are on average shorter than arbitrary sentences from MGEN documents, the length distribution contrasts with the prototypical examples in the linguistics and philosophy literature, as well as many synthetic examples in computational linguistics, that usually have less than 5 words (for example,

Text	Label 1	Label 2	Score
Puppets are fun to include too.	Particular	Unclear	0.86
First thoughts are proverbially the best; at all events, they are the bravest.	Unclear	Generic	0.96
Pumps are used to circulate the water through collectors and into your water tanks.	Particular	Generic	0.97
Players get sets by asking another player for a specific card.	Generic	Particular	0.82

Table 4: Examples of annotator disagreements with classifier scores.

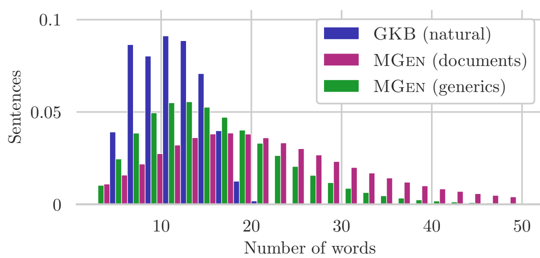

Figure 1: Sentence length distribution in the generics and documents of MGEN and natural sentences in GENERICSKB-BEST.

see Appendix F, Table F.7 and examples in the Discussion §6). Examples of sentences in MGEN with lengths from 3 to 25 words are available in Table F.9 (Appendix F).

Common words. The 50 most common words (excluding stopwords and punctuation) in MGEN also reveal interesting aspects of the use of generics (Appendix E; Table E.5).

The most common word in MGEN generics is *people*, with a big gap with respect to the second and third most common words: *also* and *cells*. In the generics of GENERICSKB-BEST, *also* is the most common word, and *water* and *one* are both more frequent than *people*, which is still fourth.

Following *people*, *women* and *children* are nouns with many occurrences, as well as terms specific to biology and medicine, such as *cells* and *patients*. The most common verb is *use* (and *used*, from passive constructions).

In contrast, we analyze the most common words in $100,000$ context documents from MGEN and find that *people* does not even appear in the top 50: it is almost 60 times less prevalent $(16,5384)$ than the most common word, which is *also* with $942,208$ appearances.

These surface statistics of the sentences in the dataset give clues as to how we use generic sentences: to generalize about *people* and to express what to *use* things for.

In biology and medicine academic domains, which are well-represented in our dataset, we find a widespread use of generic sentences, as can be seen by the high frequency of some nouns particular to those fields.

5.2 Human evaluation of MGEN

To evaluate the quality of samples in the MGEN dataset in terms of genericty we use human annotators.

We sample 300 sentences from MGEN which get annotated by two annotators by labeling the sentences as *Generic*, *Particular* (non-generic) or *Unclear*. Annotator guidelines are available in Appendix D. Examples with both annotations and the score of the ROBERTA classifier can be found in Table 4 and Table F.8 (Appendix F).

Annotators label 87.17% sentences as *Generic*, 7.5% as *Unclear* and 5.33% as *Particular*, with an 82% of inter-annotator agreement. Table 4 contains examples of disagreements. The human evaluation results suggest that, even as the annotation of generics is done automatically by a rather small model, the overall quality of the samples in MGEN is high, making it a reliable source for generic sentences in context.

5.3 Diversity

We evaluate the diversity of the MGEN dataset using three different measures: cosine similarity of sentence embeddings, distinct n-grams and distinct lemmas at subject, verb and object head positions.

Diversity from cosine similarity. Tevet and Berant (2021) introduce a transformation from pairwise sentence similarity to a diversity metric by taking an average of the similarity across possible sentence pairs (Eq. 1).

Given a corpus $\mathcal{C}$ and a 2-sentence similarity metric $m_{\text{sim}}(s_1, s_2) \in \mathbb{R}; s_1, s_2 \in \mathcal{C}$, the corre-

	diversity-from-similarity	distinct n-grams ($1M$ tokens)			head lemmas ($200k$ sentences)		
	$m_{\texttt{cossim}}$	distinct-1	distinct-2	distinct-3	Subject	Verb	Object
MGEN	$\mathbf{-7.09 \pm 0.13}$	$\mathbf{31,554}$	$\mathbf{396,923}$	$\mathbf{700,782}$	$\mathbf{18,836}$	$\mathbf{7,131}$	$\mathbf{15,935}$
GENERICSKB	-8.27 ± 0.14	$24,130$	$308,320$	$561,549$	$14,445$	$5,133$	$11,548$
GEN-A-TOMIC	-15.64 ± 0.2	$19,398$	$193,618$	$357,334$	$12,120$	$3,909$	$11,093$

Table 5: Diversity comparison of MGEN, GENERICSKB-BEST and GEN-A-TOMIC. In all scores higher is better.

sponding diversity-from-similarity metric as:

$$D_{\text{sim}}(\mathcal{C}) = -\frac{1}{\binom{|\mathcal{C}|}{2}} \sum_{s_i, s_j \in \mathcal{C}; i<j} m_{\text{sim}}(s_i, s_j) \quad (1)$$

We use as similarity function the cosine similarity ($m_{\texttt{cossim}}$) between sentence embeddings generated with NV-EMBED-V2 (Lee et al., 2024), a state-of-the-art model[4] in the Massive Text Embedding Benchmark (Muennighoff et al., 2023).

This diversity metric is computationally intractable for datasets with millions of sentences, we instead take 1000 samples of 1000 sentences each from the different datasets and report average diversity.

Diversity in distinct n-grams. We also consider an n-gram based diversity score, the distinct-n score (Li et al., 2015).

Given a corpus $\mathcal{C}$ with N_n n-grams and U_n unique n-grams. Then, the *distinct-n* score of $\mathcal{C}$ is the number of distinct n-grams (U_n) divided by the total number of words (N_1) in the corpus.

$$\text{distinct-}n_{\mathcal{C}} = \frac{U_n}{N_1} \quad (2)$$

We sample sentences from the each dataset until we reach 1 million tokens (as per the ROBERTA tokenizer). For clarity, we report the number of distinct n-grams directly, without normalizing by N_1, as all samples have the same size in total tokens.

Diversity from head lemmas. Because sentences in MGEN are naturally occurring, samples may have relative, subordinated or conjunctive clauses beyond the main bare plural generic, which could artificially inflate the n-gram count.

To have a fair comparison in this regard we introduce a score that counts the unique lemmatized verbs and head nouns in the subject and object positions. For each generic sentence, we get at most 3 lemmas, regardless of any clauses or subordinated sentences. For example, given *bees in the forests of Catalonia feed on lavender flowers, giving their honey a distinctive taste* would be reduced to 3 lemmas: *bee*, *feed* and *flower*. This way we target more directly the diversity in the generic sentences of the dataset.

We sample $200,000$ sentences from each dataset and report the total unique lemmas found.

MGEN is the most diverse generics dataset. We compare MGEN to GENERICSKB-BEST and GEN-A-TOMIC in terms of diversity by the three previous measures (Table 5). To make the comparison fair, we leave out synthetic samples from GENERICSKB-BEST, and use only the naturally occurring sentences.

In all cases, MGEN is more diverse than the comparable datasets of generics, both in lexical (distinct n-grams and head lemmas) and neural (cosine similarity) measures. This shows that the ROBERTA classifier, even if it is based on a relatively small model, is able to label a wide range of generics.

6 Discussion

In recent years, the study of generic sentences has focused on the careful consideration of a series of prototypical examples that highlight different aspects of their semantics. Some notable generics are *typhoons arise in this part of the Pacific* (Carlson, 1977b), *mosquitoes carry the West Nile virus* (Leslie, 2008), *ducks lay eggs* (Leslie et al., 2011), *humans kill themselves* (Sterken, 2015), *dobermans have floppy ears* (Nickel, 2016) and many others. Although these examples are effective at illustrating the semantics of generics, they are difficult to leverage computationally.

With the introduction of MGEN, a massive collection of naturally occurring generics in context, we open the door for new computational and corpus-level approaches to make progress in the puzzle of generics.

MGEN consists of 3 million generics and 1 million sentences explicitly quantified by 11 different quantifiers. These have been mined from a diverse pool of internet and academic documents, ensuring that many of the ways in which speakers use

[4] As of December 2024.

generics are represented.

Our analysis shows that MGEN is the more diverse of the large-scale datasets of generics, and human annotation suggests that, even as generics are automatically filtered, the quality of the examples is high.

If we take MGEN as a representative sample of generics, at least of some of the many ways in which English speakers use them, the statistics of the dataset say much about generics themselves.

The analysis of sentences in MGEN suggests that *generics are long*. They have over 16 words on average, with the most common sentence length being 15. Even if some generics in the dataset are long due to clauses and subordinate sentences, this still suggest sentences that begin with a generic express complex ideas. We also find many generics, in scientific and medical domains (Peters et al., 2024), that are not only long but contain many technical terms.

The technicality and length of many generics in MGEN contrasts with theories that link generics to "thinking-fast" or System I (Kahneman, 2011) in the dual-process theory of cognition (Leslie, 2007b; Almotahari, 2023). Combining the intuitive and unreflective use of generics, which speakers often do, with some of the long and complex sentences in MGEN is one of the open questions this dataset could help resolve.

We believe MGEN can play a role in future research on generics and quantifiers by providing examples with long context documents across a multiple sentence lengths (Appenix F; Table F.9) and topics, from academic papers to internet forums. These could disclose different ways in which speakers use generics. For example, that *people* is the most common noun suggests that generics play an important role on how humans understand each other through language.

7 Conclusion

In this work we build MGEN, a massive collection of generic and quantified sentences in context.

We mine generic sentences from ZYDA, a corpus for language model training. Our two-step pipeline first filters sentences by their syntactic features and then uses a ROBERTA-based classifier to determine genericity.

The final dataset contains over 3 million bare plural generics and 1 million quantified sentences with 11 different quantifiers. We believe MGEN is a valuable resource for future research on generic sentences.

The MGEN dataset is open-source, available at gustavocilleruelo.com/mgen.

Acknowledgements

This work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436].

The first author would like to thank Mahrad Almotahari, Dan Lassiter, Nicolas Navarre and Aina Centelles Tarrés for their proof-reading corrections and discussions.

We also thank the anonymous reviewers for their insightful comments and feedback on the manuscript.

Limitations

Data contamination. This dataset is designed as a corpus for the study of language, rather than for any evaluation of the performance of language models. The sources that conform ZYDA are commonly used in the training of language models, which means any sort of performance evaluation in this data would be compromised and should be carefully carried out.

Generics classifier. The classifier that we use to classify generics as such does only take information from the sentence itself, we do not append any context. Future versions of the pipeline could use stronger models for selection of generics from bare plural sentences.

Distribution of generics. Although MGEN has millions of generics, it may not capture the full distribution of generic sentences: it only contains bare plural generics at the beginning of the sentence. Similarly, the quantified sentences we select are within a limited range of structures.

Three main assumptions underlie the generics of this dataset: (i) bare plurals (ii) at the beginning of the sentence (iii) in English. Future work that tries to capture more holistically generics across languages should improve upon these.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference* *of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, Instantiations, and Overgeneralization: Insights into How Language Models Process Generics. *Computational Linguistics*, pages 1–60.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2618–2635, Dubrovnik, Croatia. Association for Computational Linguistics.

Mahrad Almotahari. 2022. Weak generics. *Analysis*, 82(3):405–409.

Mahrad Almotahari. 2023. Generic cognition: A neglected source of context sensitivity. *Mind and Language*.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2d2: Inductive knowledge distillation with neurologic and self-imitation. *Preprint*, arXiv:2212.09246.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *CoRR*, abs/2005.00660.

Anne Bosse. 2021. Generics: Some (non) specifics. *Synthese*, (5-6):14383–14401.

Greg N. Carlson, editor. 1977b. *Reference to Kinds in English*.

Greg N. Carlson and Francis Jeffry Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

Gregory N. Carlson. 1977. A unified analysis of the english bare plural. *Linguistics and Philosophy*, 1:413–457.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025. Generics are puzzling. can language models find the missing piece? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6571–6588, Abu Dhabi, UAE. Association for Computational Linguistics.

Ariel Cohen. 1999a. Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, 22(3):221–253.

Ariel Cohen. 1999b. *Think Generic!: The Meaning and Use of Generic Sentences*. CSLI, Stanford.

Claudia Collacciani, Giulia Rambelli, and Marianna Bolognesi. 2024. Quantifying generalizations: Exploring the divide between human and llms' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11811–11822. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Aurélie Herbelot and Eva Maria Vecchi. 2016. Many speakers, many worlds: Interannotator variations in the quantification of feature norms. *Linguistic Issues in Language Technology*, 13.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Matthew Kenney. 2023. arxiv_s2orc_parsed.

James Ravi Kirkpatrick. 2023. The dynamics of generics. *Journal of Semantics*, 40(4):523–548.

James Ravi Kirkpatrick. 2024. Are generics quantificational? *Synthese*, 204(17).

Anna Kurek-Przybilski and Adam. 2022. Generics as a paradigm: A corpus-based study of norwegian.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Sarah-Jane Leslie. 2007a. *Generics, cognition, and comprehension*. Ph.d. dissertation, Princeton University. Order No. 3256578.

Sarah-Jane Leslie. 2008. Generics: Cognition and acquisition. *Philosophical Review*, 117(1).

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.

Sarah-Jane Leslie. 2007b. Generics and the structure of the mind. *Philosophical Perspectives*, 21:375 – 403.

David Lewis. 1975. Adverbs of quantification. pages 5–20.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *CoRR*, abs/1510.03055.

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. *Preprint*, arXiv:1907.11692.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Eleonore Neufeld, Annie Bosse, Guillermo Del Pinal, and Rachel Sterken. 2025. Giving generic language another thought. *WIREs Cognitive Science*.

Anthony Nguyen. 2020. The radical account of bare plural generics. *Philosophical Studies*, 177(5):1303–1331.

Bernhard Nickel. 2016. *Between Logic and the World: An Integrated Theory of Generics*. Oxford University Press UK, Oxford, GB.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Uwe Peters, Henrik Sherling, and Benjamin Chin-Yee. 2024. Hasty generalizations and generics in medical research: A systematic review. *PLOS ONE*, 19.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Luca Soldaini and Kyle Lo. 2023. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI. ODC-By, `https://github.com/allenai/pes2o`.

Rachel Sterken. 2015. Generics in context. *Philosophers' Imprint*, 15:1–30.

Preston Stovall. 2019. Characterizing generics are material inference tickets: A proof-theoretic analysis. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Michael Henry Tessler and Noah D. Goodman. 2016. The language of generalization. *CoRR*, abs/1608.02926.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. 2024. Zyda: A 1.3t dataset for open language modeling. *Preprint*, arXiv:2406.01981.

Thomas P. Urbach and Marta Kutas. 2010. Quantifiers more or less quantify on-line: Erp evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2):158–179.

A Training and evaluation of the generics classifier

Training. We build the generics classifier by training a first iteration on generics from GENERICSKB and then refining it iteratively. We make the training set more complete by adding examples the classifier struggles on from the candidate bare plurals, thus covering difficult and corner cases. We synthetically augment this challenging datapoints with the prompts in Appendix B. Table A.1 shows the final distribution of the training dataset, which trains a classifier that reaches 0.97 F-1 score in our 3622 sentences evaluation set.

Origin	**Sentences**
GENERICSKB (generics)	2500
Synthetic non-generics	2039
Non-generics from data	310
Generics from data	61

Table A.1: Composition of the ROBERTA classifier training data.

Evaluation data. We evaluate the generics classifiers in CONGEN for positive examples and a synthetic negative examples generated with GPT-4 (OpenAI et al., 2024). We include the quantified sentences in CONGEN by removing the quantifier (*most tigers hunt rabbits* becomes *tigers hunt rabbits*). The negative (non-generic) sentences are designed to be challenging for a generics classifier (details are available in Appendix B). The final test set includes 3622 test sentences: 2873 generics and 749 non-generics.

B Synthetic adversarial non-generic bare plurals generation

We combine variations of the following prompts to generate synthetic data based on difficult examples in the data, where iterations of the generics classifier struggle. We also focus on filtering out some examples undetectable to the synthetic filtering step, such as sentences with the title section present (for example, *Introduction Transformers are function approximators*). We use some of the synthetic examples generated for the training and some for the evaluation of the classifier.

Prompt#1.
```
Task: generation of declarative sentences indicative that are not generic. The sentences generated should not be generic sentences, even if they share features with them. The following examples are non-generic sentences, or sentences that do not begin with the generic sentence.

Examples:
{ list of examples}

Based on the previous examples, generate 100 non-generic sentences using a wide range of vocabulary and basing the generated sentences on the types of syntax in the examples, and other varied
```

```
syntactic constructions similar to bare plurals, such as adding elements that make it so that the generic sentence is not at the beginning or is not grammatical. The setences cannot begin with a generic, such as "tigers have stripes" or "nerves carry messages throughout the body", but rather existentials, ungrammatical or beginning with a section title. Generate the examples in the format of a python list of strings.
```

Prompt#2. `Task: generate existential sentences that syntactically resemble bare plural generic sentences. For examples are sentences that talk about figures, equations, examples and studies in scientific articles, such as "Blue arrows indicate acceleration", "Examples of this are equations 2 and 4" or "Studies show this phenomena happens often". Can you generate 100 sentences like these in a python list sentence. Make them with varied lengths and lexically varied, and make sure they are clearly not generic, for example by referencing figure numbers etc.`

Prompt#3. `Generate 10 sentences that have a similar structure than the following example. Return the results in the format of a python list.`

```
Example: Processes are made of repetitive...
```

C Sentence Length in MGEN

The 20,000 sampled documents sampled from MGEN yield a total of 4,202,451 sentences.

Dataset	Average	Median
MGEN (generics)	16.65 ± 8.2	15
MGEN (documents)	24.75±29.3	21
GENERICSKB-BEST (natural)	9.66 ± 3.66	10

Table C.2: Average and median length across datasets.

D Annotation of MGEN

These are the instructions and examples annotators received:

- Assign the label “Generic”, “Particular” or “Unclear” to each sentence in your sheet.
- “Generic” sentences make a broad statement that applies to members of a category or group in general. For example, *Birds fly*, *German shepherds are loyal*, *Well-maintained public parks attract visitors all year-round*. Even if the group is very specific, such as *Red birds with long beaks that live in the jungle fly*, as long as it does not appear like the text refers to specific individuals in the context, label it as a generic.
- “Particular” sentences talk about a specific set of individuals or events. They usually provide information about one or a few individuals in a group: *This bird can fly*, *Dogs are in the front lawn*. These are sentences that talk about particular things in a context: *Units are in kilograms*, *Arrows indicate acceleration* would not be generics as they only make sense when refering to a specific table or plot. *German shepherds outside the house are loyal* is also not a generic, as it refers to specific german shepherds.
- In case of subsentences, focus only on the first subsentence: *Birds fly and this parrot speaks* would still count as generic even if "this parrot speaks" is not a generic since it refers to a particular parrot.
- Do not worry if you are unsure about whether a sentence is “Generic” or “Particular”. In this case, or if the sentence is grammatically incorrect, please use the “Unclear” label. Use also "Unclear" if you are not sure, you would need more context to answer or if the first words in the sentence are not a generic (for example: *In any case, birds fly*)
- For more examples, have a look at the annotated sentences in red. Thank you for your participation!

They also had the following examples:

- Tigers have stripes. *Generic*
- Tigers have stripes, they are cats and the ones we have here are violent. *Generic*
- Those tigers have stripes. *Particular*
- Tigers, which are part of the Felidae family, have stripes. *Generic*
- Tigers in this zoo are violent. *Particular*

- Tigers in zoos are violent. *Generic*
- Tigers are in the front lawn. *Particular*
- Tigers are also like this. *Generic*
- Tigers share that characteristic with lions. *Generic*

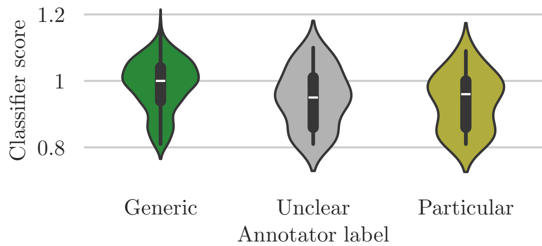

Figure D.1: Correspondence of human annotations with RoBERTa classifier scores.

E Composition of the MGen dataset

Table E.3 shows the millions of documents each component of Zyda has. Note that we only mine generics from about 23% of the dataset. The final amount of sentences in MGen by source is in Table E.4.

Finally, Table E.5 shows the top 50 common words for generics in MGen, naturally occurring sentences in GenericsKB-Best and 100,000 documents sampled from the contexts in MGen.

Source	Total Documents	Origin
RefinedWeb	$920.5M$	Internet
SlimPajama	$142.3M$	Internet
The Pile	$64.9M$	Varied
peS2o	$35.7M$	Academic
arXiv	$0.3M$	Academic

Table E.3: Information on the components of Zyda we run the generics pipeline on.

Source	Sentences
RefinedWeb	1,270,280
The Pile	1,019,687
SlimPajama	993,373
peS2o	796,334
arXiv	67,086

Table E.4: Combined statistics for MGen by source.

F Data samples

MGEN (generics)		GENERICSKB-BEST		MGEN (100k documents)	
Word	**Count**	**Word**	**Count**	**Word**	**Count**
people	200946	also	23933	also	942208
also	183012	water	20301	data	879361
cells	96700	one	18145	using	780702
used	96104	people	16598	one	767704
different	94097	many	12452	model	735504
use	92326	important	12417	used	727311
like	89778	life	11283	two	653421
one	84314	plants	10967	different	591577
make	74173	cause	10933	figure	587311
high	70107	common	10923	time	585129
many	70083	used	10715	study	584773
need	70010	body	10344	results	576442
women	68460	use	10074	may	568490
time	64141	different	10036	cells	539390
children	61270	food	9964	al.	535876
well	60362	animals	9315	however	477362
systems	60005	energy	8891	use	476105
tend	57323	human	8886	number	474336
important	56710	cells	8858	system	468788
provide	56523	form	8660	analysis	446709
work	55676	time	8478	first	445497
less	50941	children	7757	fig	438667
good	50521	women	7618	based	385968
much	48714	blood	7147	models	373924
get	47917	light	7109	high	372224
large	47588	small	7086	function	371581
small	47149	disease	6953	learning	370877
water	46181	world	6884	information	370467
way	45507	cancer	6653	case	356658
even	44487	natural	6583	set	351422
common	44330	like	6527	shown	349042
may	43538	part	6452	table	348287
patients	43443	often	6257	cell	341799
likely	43303	large	6220	new	334611
higher	43208	make	6199	given	330825
health	42758	high	6148	well	326821
help	41548	air	6017	studies	325837
men	40689	health	5982	patients	325434
system	40548	live	5889	research	321275
known	40036	two	5774	found	319645
play	39813	way	5503	could	317444
two	38604	well	5478	due	314760
human	38571	means	5464	see	312387
life	38428	occurs	5447	systems	306782
data	37663	process	5403	energy	304915
great	37612	soil	5397	thus	303428
form	37517	occur	5373	method	299352
new	37113	growth	5157	process	298258
n’t	36267	work	5145	group	290830
social	36212	system	5046	would	289965

Table E.5: Top 50 common words in generic sentences from MGEN and GENERICSKB-BEST.

Bare plural	**Source**
Solid lines are the analytical results (Eqs.	arXiv
State police report 30 year old Kira Zink was headed south ...	SlimPajama
Svp binding sites are underlined.	The Pile
COST: Entries start at $10; MORE INFO TUESDAY, DECEMBER 24...	SlimPajama
Online master's programs close on May 5th and August 19th.	SlimPajama
Tickets cost £12 (students £5, under 18s go free)...	RefinedWeb

Table F.6: Examples of existential (non-generic) bare plurals from ZYDA. Dots (...) indicate the example was truncated.

Sentences	**Source**
Horses are mammals	(Carlson, 1977)
Horses are larger than mules	(Carlson, 1977)
Elephants are easily trained	(Carlson, 1977)
Mosquitoes carry the West Nile virus	(Leslie, 2008)
Cats have whiskers	(Leslie, 2008)
Peacocks have fabulous blue tails	(Leslie, 2008)
Diamonds are valuable	(Nickel, 2016)
Elephants live in Africa or Asia	(Nickel, 2016)
Coke bottles have short necks	(Nickel, 2016)
Cabs are yellow	(Sterken, 2015)
Birds lay eggs, but mammals don't. Mammals give birth to live young.	(Sterken, 2015)
Lottery tickets are losers	(Sterken, 2015)

Table F.7: Some generics that serve as running examples in the literature.

Text	Label 1	Label 2	Score
Textbooks provide templates for proper procedure: the who, why, what, and where of the story.	Generic	Generic	0.91
Flatforms are comfy because of the uniform thickness of the heel and at the same time practical and easy to style in the morning with jeans and T-shirts and in the evening with Oversized Dresses.	Generic	Generic	0.90
Males have two sex organs, known as hemipenes, which are normally kept within the body, but are everted from his vent for mating.	Unclear	Generic	1.06
Cash crops are called commercial or commercial crops.	Generic	Generic	1.03
Oil-based primers are also very good remedies for covering staining on walls and ceilings that have oil-based paints.	Generic	Generic	1.02
Thin clients are less intelligent terminals that connect to applications hosted on a remote computer.	Unclear	Generic	1.03
Thicker greens such as romaine or bib lettuce are better for salads that will have a lot of meat or chunky vegetables.	Generic	Generic	1.07
JWs today have a similar command structure to promote uniformity rather than truth and love, in every element of a Christians life.	Generic	Generic	0.95
People realize that the best way to control their housing costs is ownership.	Generic	Generic	1.03
People who wish to argue against Spiritualism are quite sure, as a rule, that media will descend to any trickery and cheating for the sake of gain.	Generic	Generic	0.93
Red d'Anjou pears are excellent for fresh eating, poaching, cooking and all types of baking.	Generic	Generic	0.95
Powerful computing systems also require high speed access to large data storage systems.	Generic	Generic	0.95
Filipinos of Hispanic ancestry form a minority in the Philippine population.	Generic	Generic	1.06
IMTs operate in various ways.	Generic	Unclear	0.99
Weak institutions lead to weak coordination and fragmented interventions that often prove ineffective.	Generic	Generic	1.04
Ventilation flaps are used in the air ducts of heating and ventilation systems or air conditioning systems in an automobile and are usually adjusted via Bowden pull mechanisms or mechanical transmissions.	Generic	Generic	1.05
Quantum computers promise to directly simulate systems governed by quantum principles, such as molecules or materials, since the quantum bits themselves are quantum objects.	Generic	Generic	1.04
Pair bonds are monogamous and seasonal. 3–6 eggs are incubated by the female only, but the chicks are usually brooded and fed by both birds.	Generic	Generic	1.03
Puppets are fun to include too.	Particular	Unclear	0.86
Parenchyma cells are also responsible for healing in the plant - this tissue can go through cell division and regenerate when needed.	Generic	Generic	1.03
Conventional linear synchronous motors have issues of high manufacturing cost of the stator and high magnetic loss.	Generic	Generic	0.99
Traditions are a vital a part of the Italian culture and naturally, weddings have their very own.	Generic	Unclear	0.92
Calm dog breeds include Great Danes, Great Pyrenees, Basset Hounds, Shih Tzus, and Pugs.	Unclear	Unclear	0.84
First thoughts are proverbially the best; at all events, they are the bravest.	Unclear	Generic	0.96
Bursts are by definition variable, as temperature evolution due to thermonuclear burning and then cooling drives the fast increase and then slower decrease in X-ray flux.	Particular	Generic	0.97
People are under pressure to make the systems efficient, but they are expected to keep the system safe, which inevitably introduces inefficiencies.	Particular	Generic	0.91
Police officers are human beings, and many of them understand that the pressures of everyday life can sometimes lead good drivers to make bad decisions.	Generic	Generic	1.11
Self-induction habits are oft described as a compulsive behavior, with magnetic-like attraction to light sources commonly reported [9].	Generic	Generic	0.88
Gastroenterologists, infectious disease specialists, hepatologists, and even some nurse practitioners commonly manage cases of Hep C.	Unclear	Generic	1.1
Natural degradable polymers and their composites are amongst these materials.	Particular	Generic	0.84
Involving surrounding tissue structures, tonsillar tumours often infiltrate the soft palate, the base of the tongue, the lateral pharyngeal wall and medially the parapharyngeal space as well as the vascular sheath.	Generic	Unclear	0.83
Caries are understood to result from the accumulation of plaque on the teeth and the production of organic acids (plaque acids) when plaque microorganisms ferment sugars and starches in food.	Generic	Generic	1.06
Female beetles deposit their eggs singly on the legume seeds.	Generic	Generic	1.06

Table F.8: 33 examples from MGen generics with both annotations and scores.

Length	Generic	Source	Score
3	Words have power.	RefinedWeb	0.98
4	Democrats are control freaks.	The Pile	1.01
5	Children learn what they live.	The Pile	1.08
6	Ghosts represent a post-death human consciousness.	SlimPajama	1.02
7	Color and pictures are fun and vibrant.	RefinedWeb	0.82
8	More complex bytecodes trap to a software routine.	peS2o	0.85
9	Males tend to be more affected by the disease.	SlimPajama	0.99
10	Triggers cause individuals to become ineffective and produce negative energy.	The Pile	1.02
11	Professional massage therapists relieve tired muscles and alleviate pain in customers.	RefinedWeb	0.97
12	American workers produce sophisticated goods or investment opportunities at lower opportunity costs.	SlimPajama	1.06
13	Insurance companies reward property owners who personal their house totally free and obvious.	RefinedWeb	1.0
14	Alkaline phosphatases carry out hydrolase/transferase reactions on phosphate-containing substrates at a high pH optimum.	The Pile	1.0
15	Stimulants are substances that raise the levels of physiological or nervous activity in the body.	RefinedWen	1.04
16	Areas along large rivers are commonly inhabited by baldcypress, water tupelo, water elm, and bitter pecan.	The Pile	0.94
17	Sports fans are far more familiar with NBC Sports, which televises everything from Super Bowls to Olympics.	The Pile	0.96
18	Keto dieters love exogenous ketones because they help fight the keto flu and get you quickly into ketosis.	The Pile	1.07
19	Insects evolve adaptations allowing them to eat specific species of plants, while being unable to eat most other plants.	RefinedWeb	1.04
20	Extractive methods, such as lipoplasty (liposuction) or local excision, are methods whereby fat is mechanically removed from areas of interest.	The Pile	0.96
21	Factory-terminated systems are also the only viable solution to the extremely low-loss systems that are required to support high-speed optic links.	RefinedWeb	0.86
22	Small Business consultants typically develop relationships with their customers and often correspond by e-mail with their customers and return customers' phone calls.	The Pile	0.99
23	Initial parton showers interact with the medium via collisional and radiative processes that cause dissipation and redistribution of energy inside the parton shower.	peS2o	0.93
24	Green superfoods have the highest concentrations of simply digestible nutrients, fat burning compounds, nutritional vitamins and minerals to safeguard and mend your body. !	RefinedWeb	0.87
25	Punitive damages are awarded to punish a defendant for particularly egregious conduct, and to serve as a deterrent to future conduct of the same type.	The Pile	0.96

Table F.9: Examples of generics from MGEN at different sentence lengths.