Towards Linguistically Robust NLG Systems for Localized Virtual Assistants

Anonymous ACL submission

Abstract

One of the biggest challenges for localizing the natural language generation of virtual assistants like Alexa, the Google Assistant, or Siri, to many languages, is the proper handling of entities. Neural machine translation systems may translate entities literally, or introduce grammar mistakes by using the wrong inflections. The diversity of linguistic phenomena for entities across all languages is vast, yet ensuring grammatical correctness for a broad diversity of entities is critical — native speakers may find entity-related grammatical errors silly, jarring, or even offensive.

To assess linguistic robustness, we create a multilingual corpus of linguistically significant entities annotated by linguist experts. We also share a simple algorithm for how to leverage this corpus to produce linguistically diverse training and evaluation datasets. Using the Schema-Guided Dialog Dataset (DSTC8) as a test bed, we collect human translations for a subset of linguistically boosted examples to establish quality baselines for neural, template-based, and hybrid NLG systems in French (high-resource), Marathi (low-resource), and Russian (highly inflected language). We make our corpus and the derived translation-based datasets available for further research.

1 Introduction

Unlike open-domain dialog systems, the natural language interface of virtual assistants is highly task-oriented – users often interact with virtual assistants to accomplish a specific action, like finding flights, booking restaurants, buying tickets, etc.

In a task-oriented dialogue, the conversation between the user and the assistant is tracked by a dialog manager that uses a dialog state that summarizes the entire conversation up to the current turn (see, e.g., (Pieraccini, 2021)). The dialog state consists of slots and values related to the specific intents, services, and actions in question. The assistant uses the dialog state to 1) invoke external APIs with appropriate parameter values, as specified by the user over the dialog history, and 2) generate next actions to continue the dialog, for example soliciting for more information from the user or confirming the user's intent or request (Aliannejadi et al., 2021). Finally, 3) the selected dialog actions along with structured data is used to generate a new utterance to respond back to the user.

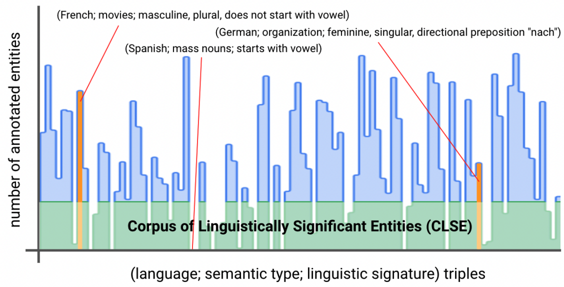

Figure 1: The corpus of linguistically significant entities (CLSE) is created by annotating a large number of entities with their linguistic properties. Entities are grouped by ($language$, $semantic\_type$, $linguistic\_signature$) triples. This results in a corpus of entities that, for a given language and semantic type, is balanced across linguistic properties. Some linguistic signatures have no annotated entities because they do not occur. For example, in Spanish, there are no mass nouns that start with a vowel.

One challenge with such systems is to design them in a way that can be robust to changes to the API's interface like renaming or adding new slot values, or integrating generalized models that can support new APIs in zero-shot settings. In this paper, we assume that we already have a system that can produce perfect English utterances – arguably a much easier task given the simplicity of English grammar – and instead focus on localization: generating fluent and grammatical output in the target language. In this setup we get the following as input: dialog state, structured data, English text, and target language. This is similar to the data-informed translation tasks of WNGT 2019 (Birch et al., 2019) and WebNLG 2020 (Zhou and Lampouras, 2020), but specifically applied to the Schema-Guided Dialog (SGD) dataset (Rastogi et al., 2020), which is task-oriented.

Ensuring grammatical correctness is arguably a basic requirement for the uniquely conversational interface of virtual assistants, in any language. Native speakers may find grammatical errors spoken by such systems silly, jarring, or even offensive, for example if the utterance refers to an entity with the wrong tense, formality, animacy, or gender (Dinan et al., 2020).

Task-oriented NLG datasets like SGD (Rastogi et al., 2020) and MultiWOZ (Budzianowski et al., 2018) are designed to be balanced with respect to the number of turns, intent, and slot usage, etc., without much focus on the linguistic properties of incoming parameter values. However, for gendered target languages like French or Italian, it could be crucial to evaluate model fluency on a test set that uses a balanced representation of masculine and feminine entities, as to observe model performance across a unbiased diversity of linguistic situations. While there has been a growing body of research on how to apply translation to data-to-text systems, the field still lacks a method for explicitly testing linguistic robustness.

A recent large-scale survey of machine translation quality using the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021) found that many of the latest machine translation models still produce various errors in fluency, accuracy, and style for inflected languages. While we have not analyzed whether the flagged grammar mistakes were specifically due to entities, this statistic is not surprising, as entities in inflected languages often have many linguistic properties that make perfect grammatical agreement by neural systems non-trivial.

Factual accuracy mistakes are also unforgivable when it comes to virtual assistants. Kale and Roy (2020) point out delexicalized models as a less error-prone alternative to lexicalized models. In the delexicalized setting, models are trained to produce output text with placeholders, which are filled in via a separate lexicalization step (usually naive string substitution). The semantic accuracy of delexicalized models tends to be far ahead of their lexicalized counterparts, especially in the presence of slot values not seen during training. However, delexicalization and other copy-based methods are more grammatically deficient in the presence of linguistic phenomena such as morphological inflection (changing surface form of a word depending on its function in a sentence; see Table 1 for an example). This makes a naive delexialization approach suboptimal for highly inflected languages (Dušek and Jurčíček, 2019).

	singular	**plural**
nominative	книга (kniga)	книги (knigi)
genitive	книги (knigi)	книг (knig)
dative	книге (knige)	книгам (knigam)
accusative	книгу (knigu)	книги (knigi)
instrumental	книгой (knigoy)	книгами (knigami)
prepositional	книге (knige)	книгах (knigakh)

Table 1: Inflections of the word "book" in Russian. Cf. "Я купил <книгу>" ("I bought a <book>"; dative) and "Моя <книга> потерялась" ("My <book> got lost; nominative).

In the search for data-to-text translation solutions that are linguistically robust and factually accurate, we create a multilingual corpus of linguistically significant entities (CLSE) annotated by linguist experts, described in Section 2. In Section 3, we discuss how we join the keys of the CLSE with the schemas described in the SGD dataset to produce linguistically diverse examples for training and evaluation. In Section 4, we describe different baseline NLG models and our experimental setup, with evaluation results discussed in Section 5.

2 Corpus of Linguistically Significant Entities (CLSE)

The Google Knowledge Graph API[1] provides access to millions of entries that describe real-world entities like people, places, and things. Each entity is a node in the graph and can be associated with any number of schema.org semantic types, such as Person, AdministrativeArea, or TouristAttraction. See (Guha et al., 2016) for more details.

We first source lexical annotations from expert linguists for a large number of entities in the knowl-

[1] https://developers.google.com/knowledge-graph

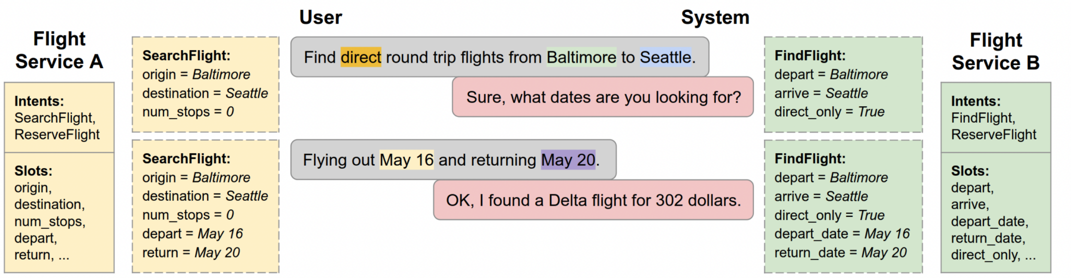

Figure 2: An example of a conversation from a schema-guided dialog (Rastogi et al., 2020). The predicted dialogue state (shown with dashed edges) for the first two user turns for an example dialogue, showing the active intent and slot assignments, with two related annotation schemas. Note that the dialogue state representation is conditioned on the schema under consideration, which is provided as input, as are the user and system utterances.

edge graph. Lexical annotations are language-specific and pertain to broader categories of linguistic properties like Animacy, Case, Classifier, Countability, Definiteness, Gender, and Number. Each language uses different linguistic properties. For example, the concept of animacy[2] is not used in the English language. Descriptions of each linguistic property class are included alongside the dataset release.

Linguistic annotations for an entity include those that are important to handle in a template-based language generation context. For instance in English, location entities have locative preposition annotations while people entities have gender annotations.[3] In other languages like French, *all* entities are annotated for grammatical gender, and entities with an article are marked depending on whether its article stays unchanged or gets merged with a preposition (like it would for common nouns).

We employ linguists who are native speakers in their corresponding language to source such annotations for popular entities. Except for the following eight low-resource languages—Bengali (bn), Gujarati (gu), Kannada (kn), Malayalam (ml), Marathi (mr), Tamil (ta), Telugu (te), and Urdu (ur)—all annotators possess at least a bachelors degree in some branch of linguistics. Linguist annotators' median age ranges from 25 to 35, and they are roughly equally split between male and female. Instead of expert linguists—or in addition to them, to cover less frequent entities—one may use data mining techniques (see, e.g., (Gutman et al., 2018)).

[2] https://en.wikipedia.org/wiki/Animacy

[3] Note that the gender annotations may not always be accurate due to changed state of the world, an annotator mistake, or a lack of standard linguistic handling for gender non-binary persons in certain languages.

We introduce the concept of a *linguistic signature*, which is a linearized string representation of an entity's linguistic attributes for a specific language. Table 2 illustrates some examples of linguistic signatures.

The maximum hypothetical number of distinct linguistic signatures for a language is the cartesian product of all linguistic features and values for that language. However, not all linguistic signatures are naturally occurring or relevant. For example, mass nouns[4] that start with a vowel do not occur in Spanish.

To obtain entities based on linguistic variation, we annotate a large number of entities for each semantic type to create a table of ($language$, $semantic\_type$, $entity\_id$, $name$, $linguistic\_signature$). We group rows in the table by ($language$, $semantic\_type$, $linguistic\_signature$) triples. The complete corpus covers 34 languages, 74 semantic types, and 222 distinct linguistic signatures.

The full Corpus of Linguistically Significant Entities (CLSE) version 1.0 is available on Github at https://clse.page.link/data.

3 Data

All of the experiments in this paper are performed on the Schema-Guided Dialog dataset from DCT8 (Rastogi et al., 2020). To emulate a data-to-text set up, we pair each utterance (text) with that utterance's acts, services, and slots. An example of such a pair can be seen in Table 3.

Owing to the lack of noun inflections in English, we can also produce a delexicalized form of the utterance: `"How would you`

[4] https://www.thoughtco.com/uncountable-noun-spanish-3079280

lang	name	signature	semantic type

fr	Suisse	`gender:FEMININE,`
		`number:SINGULAR,`
		`starts_w_vowel:0`
			Country

de	Champions League	`number:SINGULAR,`
		`gender:FEMININE,`
		`article:DEFINITE,`
		`locative_prep:PREP_IN,`
		`directional_prep:PREP_NACH`
			Event

ru	Саратовские авиалинии	`casus:NOMINATIVE,`
		`number:PLURAL,`
		`allative:PREP_K,`
		`comitative:PREP_S,`
		`topical:PREP_O,`
		`locative_prep_geo:PREP_V`
			Corporation

Table 2: Examples of CLSE linguistic signatures (truncated for conciseness).

input (structured data)	output (utterance)
service_name: "Restaurants_1" **actions:** ["OFFER_restaurant_name", "OFFER_city"] **slots:** {restaurant_name: "Bazille", city: “San Jose”}	"How would you like Bazille, which is situated in San Jose?"

Table 3: An example of structured data input and free text output for an SGD dialog utterance.

`like ${restaurant_name}, which is situated in ${city}?"` . The delexicalized form gives us an English template that allows us to substitute in other values for the placeholders to produce new realistic utterances, enabling data augmentation methods like the one described by Kale and Siddhant (2021).

3.1 Generating Linguistically Diverse Data Using the CLSE

For the problem of localization we focus on three target languages: French, Russian, and Marathi. The motivation for choosing these languages is to have a widely studied high-resource language (French), a highly inflected language (Russian), and a low-resource language (Marathi).

For entity-related slots, we source values from CLSE as an external corpus. This requires us to establish a mapping from each entity-related slot name to a semantic type. For example, *city* slots in the SGD schemas would be mapped to the City schema.org type in the CLSE corpus.

For every $(service, action, slot\_names)$ triple, we generate any number of new examples by randomly sampling slot values from the CLSE for a specific *target* language. The goal of entity resampling using the CLSE is to generate as linguistically diverse data as possible, for maximum linguistic coverage in the *target* language. It is possible that substituting entities from the target language could break the fluency of the example in the source language (English, in our case). However, we find that imperfect substitutions in the source language do not influence the quality of human translations in the target language, produced by the process described below.

3.2 Human Translations

The scope of our experiments is to focus on translation quality specifically with regard to the linguistic robustness of entities. Consequently, for the SGD dataset we only look at SYSTEM utterances, and ignore all examples from SGD's training and test data that don't use any entity slots. This results

lang	train	wt-test	SGD test	SGD dev
fr	451	233	277	187
ru	451	236	277	187
mr	451	234	277	187

Table 4: Number of items in different partitions of the SGD-NLG dataset.

in 707 $(service, action, slot\_names)$ triples from the SGD train set and 23 triples from the SGD test set.

We use the aforementioned process with three examples per $(service, action, slot\_names)$ triple to generate language-specific linguistically diverse examples in French, Russian, and Marathi[5] on the train partition of the SGD dataset. We then use two of these examples for train and one for a "within-triple" test split (wt-test).[6] Unlike the SGD dev and test splits, the wt-test split has a particular focus on the linguistics: it contains situations that were seen by the model, but with a linguistically richer set of slot values.

In addition to the train and wt-test sets, we also used a sample from the SGD dev and test sets. Each of those sets were translated to the target languages using professional translators. Table 4 has dataset statistics.

4 Experimental Setup

Below we describe our setup to test different models for their usefulness for NLG system localization.

4.1 Models for Comparison

- **nmt**: An out-of-the-box machine translation model with English plain text as input. We used the `GOOGLETRANSLATE` function of Google Sheets.[7]

- **d2t**: A data-to-text model fine-tuned on the available train set, with best checkpoint picked on the dev set. We use a pretrained mT5 xxl model (Xue et al., 2021) as a basis for our fine-tuning.[8]

- **tmpl**: Collect translated templates (delexicalized utterances) for the train set. Slot values are plugged in verbatim without any morphological inflection. Note that for most triples we have two different translations available for train (Table 4). We only use one of them (picked at random) but also report confidence intervals based on possibly picking different "base" translations as bases for templates.[9]

- **tmpl+G**: Same as above with a grammatical error correction (GEC) model applied on top of the template output. We use gT5 xxl model by Rothe et al. (2021).

Figure 3 illustrates how the input data is consumed by different baselines.

4.2 Training Details

For fine-tuning the **d2t** model we use a batch size of 64 and fine-tuned on TPU for 5’000 steps with a learning rate of 10^{-3}. We trained the models independently for each language and picked the stopping point based on the corresponding dev set. The model has 13B parameters.

4.3 Metrics

We use BLEU[10] (Papineni et al., 2002) and BLEURT[11] (Sellam et al., 2020) as our automatic metrics. For human evaluation, we assess fluency and factual accuracy.

Accuracy: Human raters are shown the original English text as well as the predicted text and are instructed to mark the predicted text as inaccurate if any information contradicts the original English text. This effectively catches errors due to hallucinations, incorrect grounding etc. Each example is rated by three raters, we take the average of the accuracy scores (1 for accurate, 0 otherwise).

[5]We use the CLSE in Hindi as a proxy due to lower linguistic annotation quality in Marathi.

[6]In cases where we were only able to obtain two examples we use one of them for train. In case of a single example, the triple is discarded completely.

[7]`https://support.google.com/docs/answer/3093331`; translations for the wt-test were obtained on 2022-01-28, for SGD test/dev: on 2022-02-02.

[8]Our early experiments with mT5 base gave substantially lower BLEU scores, suggesting that the bigger models yield stronger baselines. We use xxl models going forward.

[9]We use a simple bootstrap procedure: pick one of the two "base" translation for each triple by flipping a fair coin. The process is repeated 1000 times and the 5 and 95 percentiles are reported as the confidence interval.

[10]`https://github.com/tuetschek/e2e-metrics`

[11]BLEURT-20 checkpoint `https://github.com/google-research/bleurt`.

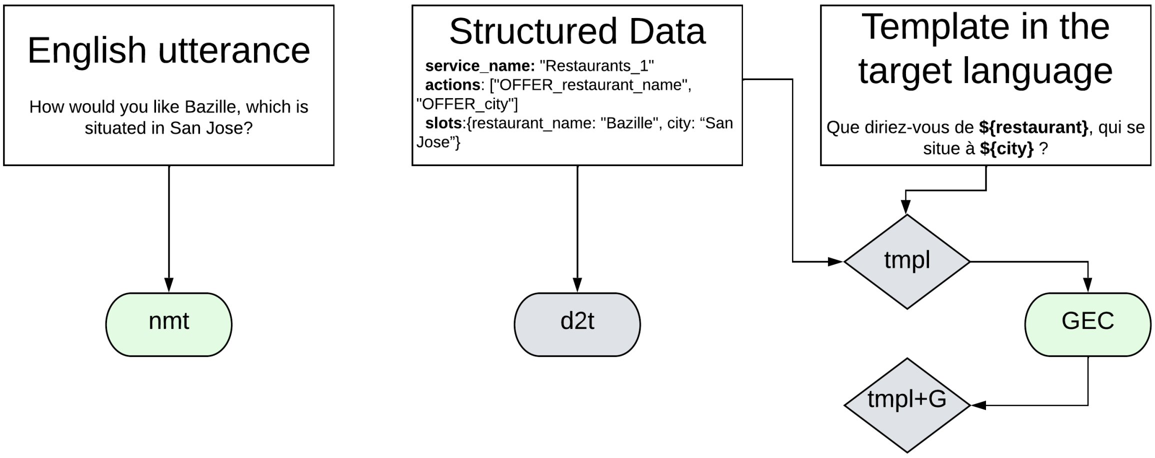

Figure 3: Input data flow for various baselines. **GEC** and **nmt** (green) are off-the-shelf models, while **d2t** (gray) is a model fine-tuned on given data. **tmpl** and **tmpl+G** are simple pipeline algorithms (rhombus shape).

Fluency: We show the predicted text to raters and ask them how grammatical the text sounds on a 1 to 5 Likert scale, with 5 being the highest score. Again, each example is rated by three raters. We average the scores across all the ratings to get the fluency score.

5 Results

We split this section into two. First, we look at the results on the unseen test set. These contain dialog situations, $(service, action, slot\_names)$ triples, which the systems did not see during training. We then move to the wt-test split, where we expect the gap to the human translations to be smaller.

5.1 Unseen Test Set

Naturally, the template-based approaches are not able to generalize to these situations, so they are not included in the results here.

Table 5 contains the summary of the results. We could verify that both baseline models are noticeably (and significantly) behind human translations in terms of fluency. In terms of accuracy, on the other hand, the **nmt** baseline is *in*significantly below the human bar for French, suggesting that we do benefit from the high-resource nature of the language. Another observation is that, while the **nmt** baseline appears to outperform **d2t** on all dimensions for a low-resource Marathi, the picture is less clear for French and Russian. There we see that the **d2t** actually scores significantly *higher* than **nmt** in terms of fluency, but *lower* in terms of accuracy. We believe that the accuracy gap could be lowered with more training data, which we leave for future work to investigate.

		BLEU	**BLEURT**	**acc.**	**fl.**
fr	d2t	0.14	0.39	0.78	4.58
	nmt	0.32	0.62	0.96$^{\blacktriangle}$	4.43$^{\blacktriangledown}$
	human	-	-	0.98	4.78$^{\blacktriangle}$
ru	d2t	0.15	0.50	0.50	4.35
	nmt	0.16	0.57	0.79$^{\blacktriangle}$	3.70$^{\blacktriangledown}$
	human	-	-	0.96$^{\blacktriangle}$	4.89$^{\blacktriangle}$
mr	d2t	0.07	0.60	0.41	3.50
	nmt	0.12	0.71	0.77$^{\blacktriangle}$	4.15$^{\blacktriangle}$
	human	-	-	0.92$^{\blacktriangle}$	4.71$^{\blacktriangle}$

Table 5: Results of the baseline models on the SGD test set (unseen triples). For human scores—accuracy (acc.) and fluency (fl.)—we also mark whether those are statistically significantly different from the *previous row* using paired t-test: $^{\blacktriangle}$ or $^{\blacktriangledown}$ denote significant difference at $p = 0.01$, $^{\triangle}$ or $^{\triangledown}$ — at $p = 0.05$ respectively.

5.2 Within-Triples Test

Table 6 contains a summary of performance of different baseline models while Table 7 contains example outputs. We see notable quality gains of the **d2t** or template-based approaches compared to the off-the-shelf **nmt** system. The generalization abilities between triples, however, are limited, possibly due to a small size of the training set.

The grammatical error correction model (**tmpl+G** baseline), appears to improve the results on top of the pure template-based **tmpl** baseline

		BLEU	BLEURT	acc.	fl.
fr	nmt	0.39	0.65	0.90	4.19
	tmpl	0.46	0.70	0.88 [0.87, 0.90]	$4.31^{\triangle}$ [4.26, 4.34]
	tmpl+G	0.47	0.71	$0.91^{\blacktriangle}$ [0.88, 0.91]	$4.48^{\blacktriangle}$ [4.43, 4.49]
	d2t	0.41	0.65	0.91	$4.64^{\blacktriangle}$
	human	-	-	0.94	4.63
ru	nmt	0.15	0.57	0.71	3.37
	d2t	0.37	0.66	0.72^{*}	$4.51^{\blacktriangle *}$
	tmpl	0.41	0.72	$0.79^{\blacktriangle}$ [0.79, 0.83]	$3.95^{\blacktriangledown}$ [3.93, 4.03]
	tmpl+G	0.44	0.74	0.79 [0.77, 0.81]	$4.21^{\blacktriangle}$ [4.17, 4.28]
	human	-	-	$0.87^{\blacktriangle}$	$4.62^{\blacktriangle}$
mr	nmt	0.15	0.69	0.72	3.73
	d2t	0.33	0.69	0.66	$4.05^{\blacktriangle}$
	tmpl	0.51	0.78	$0.83^{\blacktriangle}$ [0.80, 0.83]	$4.32^{\blacktriangle}$ [4.27, 4.34]
	tmpl+G	0.51	0.78	0.82 [0.79, 0.81]	4.29 [4.26, 4.33]
	human	-	-	$0.92^{\blacktriangle}$	$4.50^{\blacktriangle}$

Table 6: Results of the baseline models on the wt-test set (seen triples). For human scores—accuracy (acc.) and fluency (fl.)—we also mark whether those are statistically significantly different from the *previous row* using paired t-test: $^{\blacktriangle}$ or $^{\blacktriangledown}$ denote significant difference at $p = 0.01$, $^{\triangle}$ or $^{\triangledown}$ — at $p = 0.05$ respectively. The square brackets for template-based approaches denote 95% confidence intervals obtained using the bootstrap procedure described in Section 4.1.

(*) Human scores for d2t in Russian come from a different rater population and may not be directly comparable.

for high-resource languages. The gain is higher for Russian, a highly inflected language. No measurable effect is reported on Marathi, a low resource language, suggesting that the grammar error correction model itself may not be of sufficient quality. The fluency we achieve for Russian is still significantly lower than what we get from **d2t** baseline, but comes with a significantly higher accuracy (inasmuch as we can compare them given that the fluency and accuracy scores for **d2t** come from a different rater pool).

It is interesting to note that the **d2t** model appears to get higher human scores for French, but scores lower on automatic metrics. In fact, the human scores are not statistically different from that of the human baseline. Upon closer examination we see, however, that this model still frequently makes grave mistakes. One explanation for this could be that the human raters have a tolerance for hallucinations or missing facts when there are many of them presented in the same utterance. Appendix B shows examples where the automatic metrics are right to penalize the **d2t** model when it misses or introduces new facts, while the raters completely miss it. While prior work such as (Pagnoni et al., 2021; Honovich et al., 2022) studies factual consistency in English language models, further work on evaluating factual consistency of *localization* approaches is needed.

To summarize the above results, we can conclude that there is still a gap between outputs generated by our baseline systems and human translations, especially for lower-resource languages like Marathi or morphologically complex languages such as Russian. Among baseline models, the template-based approaches, when available, clearly outperform the **d2t** model for Marathi. The results of the automatic metrics in other languages suggest the same conclusion, though human scores are not always consistent with that. We attribute this inconsistency to the fact that the **d2t** system can generate seemingly good responses that, while missing or introducing facts, manage to convince human raters of their accuracy. For example, the **d2t** output in Table 7 misses multiple aspects, yet all three raters gave it a perfect accuracy score of 1 (also see Appendix B for more examples).

Even with this caveat, we can see that factual accuracy scores of the template-based approach (**tmpl+G**) is the same or higher than that of **nmt** or **d2t** baselines, suggesting that the factual accuracy of neural approaches is a fundamental weakness. At the same time, the template-based approaches still suffer from fluency issues, even with the grammatical error correction model applied. We hypothesize that the main reason is that the task of correcting mistakes in the template-based approach does not exactly map to the grammatical error correction task. There are types of mistakes we see here that humans rarely make: e.g., inserting determiners in front of popular city names like "London" or "Paris" or confusing dative for nominative (dative and accusative are more commonly confused). This suggests that a dedicated grammar model for templates may push that baseline even further.

6 Conclusion

Building a natural language generation system that can handle a broad diversity of entities with varying linguistic phenomena remains an open challenge. With the CLSE, any schema-informed NLG

Structured Data	`service_name: Hotels_2 | confirm, where_to: Amsterdam | confirm, number_of_adults: 5 | confirm, check_in_date: July 24, 2019 | confirm, check_out_date: April 22, 2021`
English	Alright. So you'd like to make a reservation for 5 in Amsterdam. You are checking in July 24, 2019 and checking out on April 22, 2021.
nmt	Bien. Donc, vous aimeriez faire une réservation pour 5 à Amsterdam. Vous vérifiez le 24 juillet 2019 et vous vérifiez le 22 avril 2021. *("check in/out" mistranslated as "verify")*
d2t	Très bien. Vous arrivez le 24 juillet 2019 et partez le 22 avril 2021. C'est bien ça? *(missing "for 5" and "Amsterdam")*
tmpl	D'accord. Vous souhaitez faire une réservation pour 5 personne à Amsterdam. Vous arrivez le 24 juillet 2019 et partez le 22 avril 2021. *("5 personne" is ungrammatical)*
tmpl+G	D'accord. Vous souhaitez faire une réservation pour 5 personnes à Amsterdam. Vous arrivez le 24 juillet 2019 et partez le 22 avril 2021.
human	D'accord. Vous souhaitez faire une réservation pour 5 personnes à Amsterdam. Vous arrivez le 24 juillet 2019 et partez le 22 avril 2021.

Table 7: Example of different baseline model outputs in French.

datasets can use techniques described in Section 3 to produce better linguistically represented data. Still, for a real virtual assistant, the space of possible entities that NLG systems will be expected to handle may be highly unconstrained, and designing solutions that are linguistically robust, and defensibly so, is an ambitious and worthwhile pursuit.

Our results in Section 5 establish an evaluation procedure probing NLG systems for their linguistic capabilities. We also evaluate four baselines using this procedure and conclude that none of them is at the human level yet. Improving upon these baselines can be approached from two sides: either 1) improving factual accuracy and reducing hallucinations of a purely neural data-to-text approach or 2) improving the quality of grammatical error correction applied to a template-based approach.

Beyond NLG applications, the dataset may also be used for the task of neural machine translation, in particular to probe the fluency aspect of the generated translations.

The full CLSE dataset is openly available on Github at `https://clse.page.link/data`.

7 Ethical Considerations

Releasing a dataset in multiple languages, including several low-resource ones, would allow to push the state of research in non-English NLG. While it is not impossible that this dataset might be used for building ML models for malicious applications, we believe it will be widely used for public good and will be a net positive contribution to society.

Crowd-sourced annotations were collected using a proprietary crowd-sourcing platform. Workers were paid at least 50% more than the minimum hourly wage. No information about the workers will be released and worker IDs are anonymized.

8 Limitations

The dataset we release and the experiments we conduct have number of limitations. Firstly, there are other linguistically significant phenomena that arise from non-entities like verbs and numbers that the CLSE does not include. Then we only cover 34 languages — small in comparison to, say, Wikipedia or Common Crawl datasets (Conneau and Lample, 2019; Xue et al., 2021) covering 100+ languages. Moreover, the quality of annotations for low-resource languages is lower due to limited linguist resources. Not only the quality, but the quantity of annotated entities varies greatly across languages, either due to some languages having fewer linguistic signatures, or annotator resource constraints. The experiments we conduct with SGD cover only a subset of CLSE in terms of languages and semantic types. The datasets we used for fine-tuning are rather small (limited by the human translation budget), and human ratings were not free of biases (e.g., humans were more forgiving for accuracy than automatic metrics). Our experiments do not include an NMT model fine-tuned on the human translations—a common domain adaptation technique (Luong and Manning, 2015; Neubig and Hu, 2018; Bapna and Firat, 2019)—in favor of a direct data-to-text model. Finally, we used the xxl versions of the models, which require significant computational resources to train and run.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *EMNLP 2021*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *EMNLP-IJCNLP 2019*.

Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh, editors. 2019. *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Hong Kong.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *EMNLP 2020*, Online. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2019. Neural generation for Czech: Data and baselines. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 563–574, Tokyo, Japan. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.

Ariel Gutman, Alexandros Andre Chaaraoui, and Pascal Fleury. 2018. Crafting a lexicon of referential expressions for nlg applications. In *LREC 2018, GLOBALEX Workshop: Lexicography & WordNets*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *NAACL 2022*, Seattle, Washington, USA. Association for Computational Linguistics.

Mihir Kale and Scott Roy. 2020. Machine translation pre-training for data-to-text generation - a case study in czech. In *INLG 2020*, pages 91–96, Dublin, Ireland. Association for Computational Linguistics.

Mihir Kale and Aditya Siddhant. 2021. Mixout: A simple yet effective data augmentation scheme for slot-filling. In *Conversational Dialogue Systems for the Next Decade*, pages 279–288. Springer.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *EMNLP 2018*, pages 875–880.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *NAACL 2021*, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Roberto Pieraccini. 2021. *AI Assistants*. MIT Press.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI 2020*, 05.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *ACL-IJCNLP 2021*, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *ACL 2020*, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL 2021*, Online. Association for Computational Linguistics.

Giulio Zhou and Gerasimos Lampouras. 2020. WebNLG challenge 2020: Language agnostic delexicalisation for multilingual RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 186–191, Dublin, Ireland. Association for Computational Linguistics.

A CLSE Dataset Statistics

Table 8 contains statistics of the CLSE dataset per language. We employ commonly used 2-letter language tags (ISO-639-1), except for "cmn-CN" for Chinese Mandarin written using the simplified script, "cmn-TW" for Chinese Mandarin written with traditional script, and "yue" for the Cantonese Chinese (written using traditional script). We do not use the Chinese macro-tag "zh" since it is important to distinguish the above locales for NLG purposes.

The complete corpus covers 34 languages, 74 semantic types, and 222 distinct linguistic signatures.

language	ar	bn	cmn-CN	cmn-TW	cs	da	de	en	es	fr	gu
# unique entities	899	721	530	529	1238	1286	2922	4076	3181	4312	798
# ling. attributes	21	6	5	6	20	23	37	32	26	26	6

language	hi	id	it	ja	jv	kn	ko	ml	mr	nl	no
# unique entities	950	705	2510	1063	55	849	885	888	924	1049	1237
# ling. attributes	18	9	41	47	2	6	9	6	6	20	23

language	pl	pt	ru	su	sv	ta	te	th	tr	ur	vi	yue
# unique entities	1606	2464	3039	29	1309	891	885	724	1262	883	612	551
# ling. attributes	30	25	31	2	19	6	6	9	10	6	8	19

Table 8: Per language statistics of the CLSE dataset. The number of annotated entities across different languages varies greatly, either due to fewer linguistic signatures (e.g., Asian languages tend to have fewer linguistic attributes overall), or annotator resource constraints.

Table 9 contains a full list of the semantic types present in the dataset.

Semantic Type	**Description**
AdministrativeArea	A geographical region, typically under the jurisdiction of a particular government.
Airline	An organization that provides flights for passengers.
Airport	An airport.
AmusementPark	An amusement park.
Article	An article, such as a news article or piece of investigative report. Newspapers and magazines have articles of many different types and this is intended to cover them all.
BodyOfWater	A body of water, such as a sea, ocean, or lake.
Book	A book.
BookSeries	A series of books. Included books can be indicated with the hasPart property.
Brand	A brand is a name used by an organization or business person for labeling a product, product group, or similar.
Bridge	A bridge.
BroadcastChannel	A unique instance of a BroadcastService on a CableOrSatelliteService lineup.
BroadcastService	A delivery service through which content is provided via broadcast over the air or online.
BusStation	A bus station.
CableOrSatelliteService	A service which provides access to media programming like TV or radio. Access may be via cable or satellite.
Cemetery	A graveyard.
City	A city or town.
CivicStructure	A public structure, such as a town hall or concert hall.
CollegeOrUniversity	A college, university, or other third-level educational institution.
Continent	One of the continents (for example, Europe or Africa).

Table 9 – continued from previous page

Semantic Type	**Description**
Corporation	Organization: A business corporation.
Country	A country.
CreativeWork	The most generic kind of creative work, including books, movies, photographs, software programs, etc.
DefenceEstablishment	A defence establishment, such as an army or navy base.
Diet	A strategy of regulating the intake of food to achieve or maintain a specific health-related goal.
EducationalOrganization	An educational organization.
Event	An event happening at a certain time and location, such as a concert, lecture, or festival. Ticketing information may be added via the offers property. Repeated events may be structured as separate Event objects.
Game	The Game type represents things which are games. These are typically rule-governed recreational activities, e.g. role-playing games in which players assume the role of characters in a fictional setting.
GovernmentOrganization	A service provided by a government organization, e.g. food stamps, veterans benefits, etc.
GovernmentService	A service provided by a government organization, e.g. food stamps, veterans benefits, etc.
Hospital	A hospital.
ItemList	A list of items of any sort—for example, Top 10 Movies About Weathermen, or Top 100 Party Songs. Not to be confused with HTML lists, which are often used only for formatting.
LakeBodyOfWater	A lake (for example, Lake Pontrachain).
LandmarksOrHistoricalBuildings	An historical landmark or building.
LocalBusiness	A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.
LodgingBusiness	A lodging business, such as a motel, hotel, or inn.
MobileApplication	A software application designed specifically to work well on a mobile device such as a telephone.
Mountain	A mountain, like Mount Whitney or Mount Everest.
Movie	A movie.
MovieSeries	A series of movies. Included movies can be indicated with the hasPart property.
MovieTheater	A movie theater.
Museum	A museum.
MusicAlbum	A collection of music tracks.
MusicComposition	A musical composition.
MusicGroup	A musical group, such as a band, an orchestra, or a choir. Can also be a solo musician.
MusicRecording	A music recording (track), usually a single song.
MusicVenue	A music venue.
Organization	An organization such as a school, NGO, corporation, club, etc.
Periodical	A publication in any medium issued in successive parts bearing numerical or chronological designations and intended, such as a magazine, scholarly journal, or newspaper to continue indefinitely.
Person	A person (alive, dead, undead, or fictional).
Place	Entities that have a somewhat fixed, physical extension.
PlaceOfWorship	Place of worship, such as a church, synagogue, or mosque.

Table 9 – continued from previous page

Semantic Type	Description
Product	Any offered product or service. For example: a pair of shoes; a concert ticket; the rental of a car; a haircut; or an episode of a TV show streamed online.
ProductModel	A datasheet or vendor specification of a product (in the sense of a prototypical description).
RadioStation	A radio station.
Restaurant	A restaurant.
RiverBodyOfWater	A river (for example, the broad majestic Shannon).
School	A school.
SingleFamilyResidence	Residence type: Single-family home.
SoftwareApplication	A software application.
SportsOrganization	Represents the collection of all sports organizations, including sports teams, governing bodies, and sports associations.
SportsTeam	Organization: Sports team.
StadiumOrArena	A stadium.
TVSeason	Season dedicated to TV broadcast and associated online delivery.
TVSeries	CreativeWorkSeries dedicated to TV broadcast and associated online delivery.
TelevisionChannel	A unique instance of a television BroadcastService on a CableOrSatelliteService lineup.
TheaterGroup	A theater group or company, for example, the Royal Shakespeare Company or Druid Theatre.
TouristAttraction	A tourist attraction. In principle any Thing can be a TouristAttraction, from a Mountain and LandmarksOrHistoricalBuildings to a LocalBusiness. This Type can be used on its own to describe a general TouristAttraction, or be used as an additionalType to add tourist attraction properties to any other type.
VideoGame	A video game is an electronic game that involves human interaction with a user interface to generate visual feedback on a video device.
VideoGameSeries	A video game series.
VisualArtwork	A work of art that is primarily visual in character.
Volcano	A volcano, like Fuji san.
Waterfall	A waterfall, like Niagara.
WebSite	A WebSite is a set of related web pages and other items typically served from a single web domain and accessible via URLs.
Zoo	A zoo.

Table 9: List of all semantic types in the CLSE with descriptions from `http://www.schema.org`. Entities can be associated with multiple semantic types. See (Guha et al., 2016) or `http://www.schema.org` for more details about the semantic types.

B Model Output Examples

We often see raters reporting perfect accuracy despite missed facts or new facts introduced by the model. This is understandable, as humans are not perfect and often miss things. Therefore, one should not completely discard automatic metrics, and use them in addition to human scores to flag potential model issues. Examples below demonstrate that, despite high human scores, such a model would be problematic to put in front of real users, as rightfully signalled by the automatic metric(s).

Table 10 exemplifies this: all results get perfect accuracy scores from the raters whereas BLEURT successfully catches the mismatch.

English reference	**d2t model prediction**	BLEURT
Satire.	Je vais lancer la lecture. C’est une satire tirée de l’album de musique techno. / [I am going to start. It’s a satire taken from **the techno music** album].	0.11
Please confirm the address of destination as 2901 East Hastings Street. Also for how many person do you want to book the cab and **what type of ride** you want?	Vous allez au 2901 East Hastings Street, c’est ça? Vous serez combien? / [You’re going to 2901 East Hastings Street, right? How many of you?]	0.16
Is your destination 2033 Camden Avenue # F3? Please tell me the number of seats you need, as well as your ride preference. You can say Pool, Luxury or something else.	Vous souhaitez une course luxe, unisexe au 2033 Camden Avenue # F3, pour combien de personnes? / [Would you like a luxury, **unisex** ride at 2033 Camden Avenue #F3, for how many people?]	0.18

Table 10: **d2t** model outputs with the lowest BLEURT scores in French. Mismatched facts are marked in bold. All three of them get perfect accuracy scores from all three raters. Human raters are imperfect, and for factual accuracy, there seems to be some tolerance for hallucinations or missing facts when there are many pieces of information presented in the same utterance.

C Datasheet

Datasheets for Datasets "document [the dataset] motivation, composition, collection process, recommended uses, and so on. [They] have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks."

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

CLSE was created for training, testing, and evaluating NLG systems in multiple languages, including several low-resource ones. It allows to do sampling and slicing by language, semantic type, or linguistic phenomena.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Anonymized for submission.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Anonymized for submission.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Entities of semantic types detailed in Appendix A.

How many instances are there in total (of each type, if appropriate)?

80'893 language entries (13'649 unique entities).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset represents a sample of all entities found in the Knowledge Graph. For each language and semantic type, the sample is meant to limit over-representation of entities with common linguistic attributes (see Figure 1).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

See Table 2 for examples.

Is there a label or target associated with each instance? If so, please provide a description.

No.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Certain linguistic attributes may not be annotated for some languages due to limited language support.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No, except for the same entity—identified by its ID—appearing for multiple languages as a separate row.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Surface forms (entity names) and linguistic annotations were created by humans and therefore may be inaccurate or incomplete.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. Entity IDs refer to the Google Knowledge Graph API, but this is as an implementation detail (API stability does not affect the usefulness of the dataset).

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, to the best of our knowledge.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Some entities in the dataset are of semantic type "Person."

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes, people with a Knowledge Graph entry can be uniquely identified.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, to the best of our knowledge.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was curated by linguists. See Section 2 for more details.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was curated in spreadsheets and text files and, as a rule, reviewed by another linguist.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Exact details of the sampling procedure cannot be disclosed at the moment to preserve anonymity and to comply with internal policies of the authors' organizations.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Contractors. Each contract is reviewed, approved, and executed according to the strict company policies.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The bulk of linguistic data was collected over the years 2020 and 2021. Semantic type associations were retrieved from the Google Knowledge Graph API on 2022-05-18.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes. The dataset description was improved and this datasheet was created as an outcome.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Some entities in the dataset are of semantic type "Person." These are limited to individuals (alive, dead, or fictional) who are popular enough to have a Knowledge Graph entry.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

No.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

N/A.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, for experiments in Section 4.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?

E.g., for balancing machine translation data.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The linguistic attributes are provided "as is" and may be innacurate or incomplete.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset should not be used to infer non-linguistic properties of entities. In particular, the linguistic attributes are not appropriate proxy data to infer a person's aliveness or gender.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

As a CSV file retrievable from `https://clse.page.link/data`.

When will the dataset be distributed?

Upon acceptance of the publication.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Yes, CC-BY license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No, to the best of our knowledge.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, to the best of our knowledge.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of this publication.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Yes, by email or any other contact point provided at `https://clse.page.link/data`.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

No updates are planned at the moment. If any is made, it will be communicated at `https://clse.page.link/data`.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Please, contact the dataset mainteners using the contact information above.